ISSN 2074-9856 (print)
ISSN 2074-9864 (online)

VIEWZ

VIENNA ENGLISH WORKING PAPERS

VOLUME 18 **NUMBER 1** **JUNE, 2009**

INTERNET EDITION AVAILABLE AT:
HTTP://ANGLISTIK.UNIVIE.AC.AT/VIEWS/

CONTENTS

LETTER FROM THE EDITORS

Dear Readers,

Just as the semester ends and before the holiday season starts, this issue of VIEWZ comes filled with three thematically different contributions which cater to a wide range of linguistic interests. *Testing*, *that* and *th* are among the phenomena discussed by the authors in the current issue.

Bringing together teachers' practical needs and assessment and testing theory, Armin Berger's article explores the process of developing a rating scale for accessing students' speaking ability at the end of a university course on practical phonetics and oral communications skills. Reconciling the local

institutional context with considerations of external validity, Berger employs a Rasch model analysis in order to test the validity and applicability of a rating scale developed by the local team of lecturers and hence presents a prime example for research-informed language teaching and assessment.

Gunther Kaltenböck's contribution presents a rich and thorough investigation of complementizer use in spoken object clauses. Focusing on initial epistemic clauses such as *I think*, Kaltenböck discusses the ambiguous and indeterminate syntactic status of these clauses, stressing the importance of contextual realisation. An in-depth prosodic analysis carried out on ICE-GB confirms Kaltenböck's initial hypothesis that *that* is largely used as a filler and, in spoken language, seems to have lost much of its subordinating syntactic function.

The third contribution of this issue, by Ruth Osimk, investigates intelligibility in English as a lingua franca (ELF) in an experimental setting. Adopting a psycholinguistic approach, Osimk builds on the findings of a previous (but methodologically different) study on the phonology of ELF and puts some hypotheses proposed in this study to the test. Focusing on the realization of three features (aspiration, interdental fricative, and /r/), she tests intelligibility of accented variations of these features employing the dictation method and thus corroborates the findings of the previous study with regard to two of the three tested ELF features.

We hope you will enjoy the range of stimulating papers in this VIEWS summer issue (on the beach or during a quiet hour at the office...) and, as always, we would be happy to include your comments in the form of a reply in our next issue.

We wish you a sunny and relaxing summer!

THE EDITORS

Testing speaking: developing a rating scale for advanced learners of English

*Armin Berger, Vienna**

1. Introduction

Information about a student's L2 speaking ability is not only useful but often necessary in many situations. Without this information it may be difficult to see how rational educational decisions such as, for instance, planning a speaking lesson, placing students in ability groups or measuring achievement, can be reached. Assessing speaking is a challenge, however, because it involves a rater making judgements about a person's speaking performance. Unlike reading or listening skills, which can be assessed by discrete items scored dichotomously as correct or incorrect, speaking skills are usually assessed in a communicative situation, in which an extended sample of speech is elicited from the test taker and judged by one or more raters. It is easy to see how factors other than the test taker's speaking ability can influence the judgements. However, the problem of subjectivity in the rating process can be minimized by establishing a clear rating procedure and a framework for making judgements. Depending on the purpose, such a framework may take the form of an analytic rating scale. This article outlines the design process of such a rating scale intended for the assessment of spoken English at tertiary level.

With the present article I address the issue of assessing advanced oral communication skills at the Department of English and American Studies at the University of Vienna, thereby satisfying the need to establish clear exam specifications. At the same time I aim to narrow the gap between the well-established assessment procedures for writing and reading skills at this department[1] and the assessment of speaking, which has generally received

* The author's e-mail for correspondence: armin.berger@univie.ac.at.

1 Such as, for example, the Common Final Test (CFT), which is a standardized test of writing and reading that every student has to take after two semesters of Integrated Language and Study Skills (ILSS 1 and 2).

less attention. This paper thus responds to the need to professionalise the assessment of oral language skills in educational settings and to increase awareness of the work involved in ensuring minimum quality of assessment instruments. Moreover, I hope to excite interest in the subject of testing and assessing speaking and language in general, for a lot more (context-specific) research is needed. I will start by providing some background information on the institutional context, the assessment of speaking, and general characteristics of the scale in the first four sections. Section five explains the concept of validity, while sections six and seven outline the scale development process and basic methodologies employed in this study, respectively. After a more detailed description of the results in section eight, I will discuss in section nine what I consider key issues arising from the data obtained when using the scale in the first live examinations and conclude with fundamental research questions that need to be addressed in the future by more comprehensive research projects. I should like to add that the rating instrument presented here, as ideally any other rating scale, is work in progress and subject to constant change and improvement. Much as this article depicts ongoing research efforts rather than final results, it accomplishes the desired aim of investigating and making known some fundamental psychometric properties of oral examinations in language departments at Austrian universities. It is intended as an important contribution to the advancement of language testing practice at tertiary level.

2. Institutional background

The Klagenfurt Language Testing Centre's website states that "[a]lthough language competence is being assessed in Austria at secondary and tertiary level in the educational system, professionalism in the current practice is largely missing" (LTC 2009). Indeed, language testing has yet a long way to go in Austria. Until recently, testing in the language programmes for students majoring in English at Austrian English departments has been a largely independent and isolated endeavour of each individual teacher. Although there are standardized language programmes with common course curricula, lecturers generally design their own instruments to test students' achievement of specific course-related objectives. Test content, format as well as assessment criteria vary among teachers, who seem to rely almost exclusively on their own testing experience. At best, there are some common guidelines regarding examination procedures such as a double marking policy, but specifications for these examinations rarely exist. The test constructs are, if at all, vaguely defined, the tasks do not always elicit the required information to

make sound inferences about a student's ability, and rating scales are often poorly constructed. That is, current testing practice at Austrian schools and university departments lacks validity in many respects.

With a general movement towards more transparency in educational systems, the demand for international comparability as to language proficiency and the resulting advent of the Common European Framework of Reference for Languages (CEFR), the testing scene has begun to change over the past decade. The ideas and resources set out in the CEFR have sensitized teachers, course designers, curriculum developers, and language testers to the lack of professionalism in the field of language testing and the critical need for action. Remedial measures in Austria include, for example, the foundation of the Language Testing Centre at Klagenfurt University and its activities, notably the Austrian University English Language Teaching and Testing (ELTT) initiative, which promotes concerted action to professionalize language assessment and certification practices at Austrian university English departments.[2] With such measures language testing ceases to be an isolated, solitary activity of individual teachers but becomes a group endeavour, in which language teachers cooperate.

Structural changes have also led to a growing demand for professionalism in language testing. Many curricula at Austrian universities have already been converted into separate bachelor and master programmes. Such major restructuring of the system raises the question of what a BA graduate in English language and literature should be able to know and do in terms of language competence. While there has been some work to answer this question with respect to listening, reading and writing skills,[3] the nature of speaking ability at tertiary level and the question of measuring it have yet to be addressed in theory and practice. The present study can be seen in this light. It is part of an ongoing process to consider issues surrounding the nature of speaking and testing a foreign language at tertiary level, attempting to fill some serious gaps in language testing practice with respect to validity and thus contributing to professionalism at Austrian language departments.

---

[2] Cf. http://www.uni-klu.ac.at/ltc/inhalt/430.htm.

[3] For example, the Austrian University ELTT Group, a working group consisting of applied linguists and university language teachers of the Universities of Graz, Klagenfurt, Salzburg and Vienna, have established an analytic rating scale for *Writing* as well as a set of benchmarked performances.

3. Assessing spoken language

The ability to speak proficiently in a foreign language and to perform different tasks for various purposes in a number of communicative situations is highly valued; yet speaking is the skill that has long been neglected in language testing research and practice. Consequently, the theory and practice of testing speaking in a second language is the youngest sub-field of language testing (Fulcher 2003: 1). There are several reasons for which Lado's (1961: 239) observation that "testing the ability to speak a foreign language is perhaps the least developed and the least practised in the language testing field" still holds true today. Firstly, speaking is a language skill difficult to assess reliably. In the test situation spoken discourse elicited by some test task is heard by a human judge who then refers to a rating scale in order to select a score that represents the candidate's ability. It is easy to see that such performance testing brings with it "potential variability in tasks and rater judgements, as sources of measurement error" (Bachman *et al*. 1995: 239). Research into rater performance investigating such variability was carried out, for example, by McNamara (1996), McNamara and Lumley (1997), O'Sullivan (2000) and Wigglesworth (1993). Secondly, construct-irrelevant facets might have an impact on the candidate's speaking performance and scores to a greater extent than in test situations assessing other skills. The nature of the interaction, the test methods, the topics, the interlocutor effect, and test taker characteristics account for some of the variability in speaking test scores (Berry 2007; Brindley 1991; Brown 2003, 2005; Kunnan 1995; O'Sullivan 2006; Shohamy 1988, 1994). Thirdly, many difficulties of assessing speaking boil down to the question 'What is speaking?'. Lado (1961: 239) argued that speaking was neglected because of "a clear lack of understanding of what constitutes speaking ability or oral production". What exactly does it mean to be able to speak, what exactly is being measured in speaking tests, or how can the construct of speaking be defined? Indeed, much of the research in language testing is concerned with the ongoing challenge of construct definition. In addition to theoretical issues, complex logistics and practical constraints make tests of spoken language more difficult to administer and research than tests of other skills. In summary, one might argue with Fulcher (1997: 75) that speaking tests are particularly problematic in terms of reliability, validity, practicality and generalisability.

All these difficulties have led Hughes (2002: 75) to raise the "question of the extent to which the characteristics of natural spoken discourse can ever lend themselves to existing assessment paradigms". There seem to be so many complexities and competing factors influencing the test scores that the assessment of speaking has tended to focus on the more quantifiable aspects

of speaking such as pronunciation or the number of grammatical errors. The question then arises whether the test construct adequately reflects the nature of oral proficiency or whether the test is still a test of speaking rather than a test of more general language proficiency measured by structural complexity and accuracy. Asked differently, is 'speaking' still the real focus of the test?

Issues surrounding the nature of oral proficiency, questions about how to best elicit it, and attempts to find effective ways to optimise the evaluation of oral performances have motivated much research in this area. While many aspects of testing speaking remain obscure, "it is important to recognise the great improvements in the area that have been made over the last few decades" (O'Sullivan 2008: 1). Despite the problems surrounding the testing of speaking, there seems to be agreement that there are ways of overcoming or at least addressing some of these problems by careful development of the testing procedures, including the careful construction of the tasks to elicit and the tools to evaluate speech as well as continuous training of raters to ensure the quality of their ratings. The following sections will describe in some more detail the development process of a rating scale used to assess the speaking ability of students after a two-semester speaking course at the Department of English at the University of Vienna.

4. The context of the PPOCS 2 rating scale

The Department of English and American Studies at the University of Vienna recognized a need to extend the speaking component of the language competence programme and developed a new course to complement the existing Practical Phonetics and Oral Communications Skills (PPOCS) syllabus. As the only speaking course, PPOCS was considered insufficient to meet the great demand for oral language proficiency and to allow for the fact that students need more time for the improvement and consolidation of their spoken communication skills. The department also acknowledged the need for extended systematic training in aspects of oral communication other than pronunciation training and accent improvement, on which the former PPOCS course heavily focused, as well as the need to redress the balance between language skills. Since the previous curriculum concentrated primarily on reading and writing, a new compulsory speaking module was developed to give greater weight to the oral component of the language.

As a consequence, the previous PPOCS course was extended in the new BA curriculum to include a greater focus on the communicative and interactive aspect of speaking. Two complementary speaking courses (PPOCS 1 and PPOCS 2) now constitute the new speaking programme. Whereas

course one tends to place emphasis on practical phonetics and the mechanical aspects of the spoken language, including pronunciation theory and training at both the segmental and suprasegmental levels, course two has a stronger focus on the sociolinguistic, pragmatic and strategic aspects of speaking. Building on the skills and knowledge featured in PPOCS 1, the new PPOCS 2 course concentrates on formal presentation and interactive speaking skills, aiming to educate expert users of spoken English in its productive and interactive form, in various stylistic, contextual, social and geographical forms of spoken English. It covers distinctive features of spoken language and provides the opportunity to practise the effective use of intonation, voice, turn-taking devices and lexico-grammatical means to interact successfully in conversation and discussion.[4]

In addition to the new course syllabus, a new assessment system had to be devised. Considering the main aims and objectives, the course developers agreed that a short obligatory in-class presentation followed by interaction, portfolio work on practical phonetics, and a final oral exam would both satisfy the legal assessment requirements and reflect the core contents of the course. Since the final oral exam was seen as the major instrument for assessing the students' oral proficiency, clear examination procedures had to be developed. As set out in the exam specifications, which include information on the tasks, administration and rating process, the final exam consists of two parts: an individual five-minute presentation, which should be a condensed and improved version of the students' in-class presentation, and a fifteen-minute spoken group interaction of four students in the form of an unrehearsed role-play, in which the candidates discuss a controversial topic from distinctly different points of view to achieve a clearly specified purpose such as finding a consensus, deciding on a plan for action or solving a problem. Both the individual presentation and the group discussion are designed to elicit extended speech samples or production responses on the basis of which meaningful inferences about the students' proficiency can be made. The two parts of the exam reflect the increasingly recognised fact that valid assessment requires the sampling of a range of relevant types of discourse in a range of task types that will allow inferences to be made from scores to constructs (Fulcher 2003: 86). Two raters assess the students' presentations and interactions separately and independently and average their scores to arrive at a final grade.

---

4 For more details cf. PPOCS course description available at http://online.univie.ac.at/vlvz?kapitel=1201&semester=S2009.

5. The nature of the scale

An essential constituent of this new assessment system is a rating scale that reflects the assessment criteria and describes the various levels of performance. In the PPOCS 2 exam situation, as in any other oral language testing, raters are interested in how well a student can speak the language being tested. In order to assess the quality of a student's oral proficiency, the speech samples produced in the formal presentation and group discussion tasks mentioned above are rated. However, in contrast to limited production responses, which can be readily assessed by a dichotomous scale as either right or wrong, extended production responses cannot be classified in this binary way. Rather, raters *judge* the quality of the response in terms of levels of ability by means of a multi-level rating scale, which is defined by Davies *et al*. (1999: 153-4) as

> *a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency … The levels or bands are commonly characterised in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency and cohesion) … Scales are descriptions of groups of typically occurring behaviours; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose.*

Fulcher (2003: 89) points out that this definition of a rating scale as "an operational definition of a linguistic construct" is based on the assumption that "the rating scale will be used to (a) score speech samples, and (b) guide test developers in the selection of tasks for tests". Indeed, the scale developed for PPOCS 2 is intended to enable university teachers to describe students' performances at successive bands of ability that are meaningful to those involved. It is, in fact, the primary function of the scale to help raters consistently make informed judgements about the quality of a student's performance. In other words, the PPOCS 2 rating scale is *assessor-oriented* as it provides guidance for assessors who are rating performances. It is a common standard for different raters, ensuring reliability and validity.[5]

[5] According to Alderson (1991) and Pollitt and Murray (1996), other purposes of rating scales include *user-orientation*, with a reporting function, *constructor-orientation* guiding the construction of tests at appropriate levels, and *diagnosis-orientation* for feedback purposes. Alderson argues that these different purposes should not be confused, because one rating scale is rarely appropriate for several functions. Therefore, it is important to determine the primary aim of the rating scale and develop it according to its

The PPOCS 2 speaking scale can be seen as resting on three major assumptions. The first one classifies the scale as an ability-based scale and refers to the notion that speaking ability is not a single unitary ability, but consists of multiple components. Anyone who wants to speak a second language must be able to use some of the grammar and vocabulary of the language, and master its specific sound system. Learners must conceptualize, formulate, articulate, monitor and, if necessary, repair their speech. They need to be able to speak with some degree of accuracy and fluency if they want their utterances to be considered acceptable. Furthermore, in interactive activities the language user is both listener and speaker so as to negotiate meaning and construct discourse conjointly. During interaction, reception and production strategies are constantly employed. Yet another class of abilities concerns the use of various cognitive and collaborative strategies to manage co-operation and interaction. All these components make speaking a complex multi-ability matter. A major issue of the scale development process was to address the question of which of these abilities were to be included in the construct definition that forms the basis of the scale. The second assumption, which follows from the first, refers to the analytic nature of the scale. The different components of speaking ability require separate analytic ratings for each of the specific components in the construct definition as opposed to one overall score of a holistic scale. In other words, an analytic rating scale contains a number of criteria, each of which has descriptors at the different levels of the scale. The third assumption is that the scale is criterion-referenced. The scale is defined operationally in terms of criterion levels of ability. Whereas norm-referenced assessment ranks test takers in relation to their peers, criterion-referencing assesses the learners purely in terms of their ability, irrespective of other test takers. Such criterion-reference scales allow the tester to make inferences about a learner's ability, and not just the quality of an individual's performance relative to other individuals. Summing up, all these assumptions qualify the PPOCS 2 speaking scale as a criterion-referenced ability-based analytic scale.

6. Validity

Such scales offering descriptions of a learner's proficiency at successive levels of ability have become very popular in language testing. Raters quickly embrace these scales and learn to use them quite successfully for their specific

specific context, rather than adopt available rating scales designed for some other purpose in some other context.

purposes. However, just because a scale is used with some efficiency does not automatically mean that the inferences drawn from that scale are valid – "there is no guarantee that the description of proficiency offered in a scale is accurate, valid or balanced" (North and Schneider 1998: 219). Since the effective use of a scale does not necessarily entail validity, the need for validity studies is well established and recognized (Butler and Stevens 1998; Matthews 1990; McKay 2000; McNamara 1996; Shohamy 1995). By providing validity evidence, testers can make sure that the inferences drawn from the scale offer an accurate picture of the underlying abilities or constructs they want to measure.

However, just as undisputed as the need for validation is, as diverse are the interpretations of the concept of validity. Earlier notions of validity were concerned with the question of "whether a test really measures what it purports to measure" (Kelley 1927: 14; cf. also Cronbach 1971; Henning 1987; Lado 1961). From this perspective, validity is regarded as a characteristic of the actual test. While some writers find such a general approach still useful (Davies 1990, Hatch and Lazaraton 1997), Messick (1989, 1996) argues that the traditional conception of validity is incomplete especially because it does not take into account evidence of the implications of score meaning or the social consequences of score use. Validity is not a quality of tests or test scores, but a quality of interpretations and uses of assessment results. Instead of speaking of the validity of a particular test or of the scores of a particular test it is more accurate to speak of the validity of the uses of a test score, or of test scores as valid indications of a specific ability. Messick sees validity as multifaceted and calls for different types of evidence to substantiate any inferences drawn from the scores on a test:

> *Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as the potential – consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility. (Messick 1995: 742)*

For the purposes of this study, a restricted definition of validity is used. According to this definition, there is evidence for the construct validity of the variable in question, when Rasch analysis as detailed below shows little misfit. Such an understanding of validity, which will be explicated in section eight, is common in the context of Rasch analysis (Tyndall and Kenyon 1996).

Extensive validation studies for the PPOCS 2 scale have yet to be carried out. The following sections summarize a few preliminary measures taken to

look for minimum validity evidence. At a minimum level, all test development activities need to be documented. This documentation, however, would not withstand scrutiny without further analysis of the scores obtained from the rating scale. The remaining sections of this report provide this documentation and data analysis for the PPOCS 2 scale in search of validity evidence.

7. The scale development process

The development process of the PPOCS 2 rating scale was divided into four stages: Stage one was an intuitive phase, in which a first draft of the scale was created. For this purpose, existing speaking scales, the relevant CEFR illustrative descriptors, and language teachers' expertise were consulted to find rating criteria and formulate level descriptors. Secondly, the first draft was edited in a qualitative stage of vetting and modifying while obtaining expert judgements from experienced PPOCS teachers. The third stage included piloting the scale in trial runs to see how well the rating scale functioned. The feedback obtained from this trial period was used to modify the rating scale in the final editing stage.[6]

The methods used to develop the first draft of the scale can be labelled as intuitive expert or committee methods (CEFR: 208). In testing it is very common that a group of expert teachers create their own context-specific test instruments, which are then discussed and commented on by a larger group of consultants. In fact, most existing rating scales for small- or medium-scale testing situations have been developed by means of intuitive methods, since these are relatively time- and cost-effective compared to other data-based methodologies. Intuitive methods are not empirically-driven by any structured data collection and analysis but rely on the principled interpretation of experience, hence 'intuitive'.

In this initial stage, the PPOCS lecturer team agreed to rate the two sections of the exam, i.e. the formal presentation and the interaction, separately and split the rating scale in two parts. The two parts are, in fact, two separate but overlapping scales that both reflect the dual focus of the new PPOCS 2 course and accord with the two CEFR categories for spoken communicative activities: production with prepared, long turns, and

6 I would like to thank Thomas Martinek and all other members of the PPOCS lecturer team, Harriet Anderson, Meta Gartner-Schwarz, Katharina Jurovsky, Gunther Kaltenböck, Sophie-Francis Kidd, Amy Krois-Lindner, Christina Laurer, Karin Richter, and Andreas Weißenbäck, for their valuable contributions to the scale development process.

interaction with spontaneous, short turns. In a series of meetings the PPOCS lecturer team compiled and discussed various components characteristic of speaking proficiency at the intended level, which were then reorganized under the following three labels: *lexico-grammatical resources*, *fluency*, and *delivery*. Whereas these three criteria are used to assess both the presentations and the interactions, two more criteria, one labelled *relevance, development and organisation of ideas* and the other one *interaction*, refer to presentations and interactions, respectively. Eventually, the first intuitive phase resulted in two sets of analytic rating criteria, each of which consists of four categories, which in turn are comprised of five level bands. While the first draft version contained four bands, it was later agreed to add one more level to the scale, mainly because teachers in Austria are used to a five-point grading system and computing an overall grade between one and five would be more straightforward. Moreover, the PPOCS team agreed that the targeted level would range between C1 and C2 of the CEFR. Thus, the rating scale was designed to stretch the CEFR levels C1 and C2 into four pass levels (*C2*, *C2-*, *C1+*, *C1*), whereas band five would indicate a level *below C1* and failure in the PPOCS 2 exam. The two extreme points of the pass range, i.e. bands *C2* and *C1*, were put into words; the two middle bands of the scale and band five indicating non-pass were left unworded (North 2003: 83).

A smaller team of PPOCS lecturers then developed a first draft of the wording of the descriptors. The starting point was to consider the main contents of the course syllabus and descriptors of existing scales, most notably the CEFR speaking scales. The most relevant CEFR descriptors used to formulate the PPOCS scale were taken from the bank of 'illustrative descriptors'[7]. In addition, other holistic as well as analytic rating scales informed the wording process.[8] The information obtained from the CEFR, other scales, the course syllabus, and the group discussions on rating criteria was distilled into draft descriptors, which were subsequently submitted to the lecturer team for a process referred to as vetting. Some preliminary decisions needed to be taken as to which descriptors should go forward for detailed

[7] These include the 'Common Reference Level: Qualitative Aspects of Spoken Language Use', 'Overall Oral Production', 'Sustained Monologue', 'Addressing Audiences', 'Overall Spoken Interaction', 'Formal Discussions and Meetings', 'Vocabulary Range', 'Vocabulary Control', 'Grammatical Accuracy', and 'Phonological Control'.

[8] The Revised ACTFL Guidelines 1999, the Test of Spoken English scale and the analytic rating scale of the Examinations Reform Teacher Support Project of the British Council Hungary were used for consultation.

editing and trialling, and which ones should be revised or redrafted immediately.

The following phase of trialling involved trying out the draft scale on a representative sample of the test-taking group to gather information about the usefulness of the scale. The presentation scale and the interaction scale were trialled separately on two different occasions. A number of students were asked to volunteer in mock exams, which were videotaped and rated independently. Since the number of ratings obtained in these trial runs was too small to carry out any quantitative analyses, it was the feedback from raters that provided the most relevant information about the usefulness of the scales at this stage. This feedback was used in a further attempt at honing the descriptors, and final adjustments and corrections were made in the light of it. The result of this phase was a final set of descriptors, which, by way of illustration, stipulate that candidates at the *C2* level "show great flexibility formulating ideas in differing linguistic forms to convey finer shades of meaning precisely" (*lexico-grammatical resources*), "elaborate all salient points in the prompt in adequate detail with examples and ideas of relevance" (*relevance, development and organisation of ideas*), and "contribute ideas of relevance to the joint discourse and display great flexibility in responding to others, e.g. by framing the issue, establishing a line of approach, proposing and evaluating, recapping, summarising, etc." (*interaction*).

8. Validation and analysis

Quantitative analyses were carried out *post hoc*. Even though great care had been taken to develop effective and unambiguous rating scales, empirical analyses were carried out to examine the quality of the measures and the utility of the rating scales in yielding interpretable results. This section reports on the analysis of the data obtained from the first live tests employing the new scales. Such investigation sheds light on the influence of both the number and the labelling of the categories. The Rasch model provides an appropriate framework for such analysis.

The basic Rasch model is a one-parameter model in Item Response Theory (IRT), which allows the calibration of items and persons on a linear scale. In other words, the analysis models the expected behaviour of individual candidates on a test item by estimating person ability and item difficulty. It facilitates two different facets of the data to be analysed on the

same scale yet independently of one another.[9] While the basic Rasch model calculates person ability and item difficulty, the multi-faceted Rasch analysis is an extension of the simple Rasch model and takes into account additional variables (facets) of the test situation, such as raters or rating category, indicating rater severity and rating category difficulty, respectively. Person ability is then estimated while the effects of other variables are taken into account. Furthermore, the analysis shows 'misfitting' elements within a facet. The fit statistics identify unsystematic elements as, for example, raters who are unsystematically inconsistent in their judgements or rating criteria that are unsystematically difficult across all observations. All this information is useful for the construction of rating scales, rater training, and test design.[10]

Traditionally, Rasch analysis has been used in the testing of speaking to analyse existing rating scales *post hoc* to address validity concerns (Stansfield and Kenyon, 1992b). More recent approaches build validity questions into the design process and therefore use Rasch analysis when developing rating scales (Fulcher 1993; McNamara 1996; Milanovic *et al*. 1996; North 1995, 2000). In the present study, the rating scale was developed according to intuitive methods, and Rasch analysis is used to examine whether the scale is operating as intended.

The edited version of the scale was first used live in the testing period in February 2009. A total number of 36 students took the new PPOCS 2 exam, and two raters assessed their performance on each of the four dimensions for both presentation and interaction.[11] A common ground in terms of scale interpretation was established when the two raters participated in the qualitative phase of the scale construction process. Each speech sample was rated independently by the two raters, who assigned a level between one and five for each category based on the rating scale, where *one* indicates the most competent (*C2*) and *five* the least competent performance (*below C1*). This yielded a total number of 576 data points, which were analyzed by means of FACETS (Linacre 2008), a computer program used for multi-faceted Rasch analysis.

In this study, four facets were examined to see how the rating scale functioned: raters, candidates, types of discourse, and rating criteria. To investigate how well the PPOCS 2 rating scales operated, two pertinent questions guided the multi-faceted Rasch analysis: a) How effective are the

---

9 The interested reader is referred to Baker (1997), who provides an accessible introduction to IRT and points out the advantages over Classical Test Theory (CTT).

10 See McNamara (1996) for more on multi-faceted Rasch measurement.

11 All performances were video-recorded for rater training and research purposes.

analytic rating criteria? That is, do raters use all level bands for the test population? b) How well do the facets examined in this analysis fit into a multi-faceted Rasch model of speaking performance? Asked differently, can the band level descriptors be distinguished adequately and thus performances rated systematically? This question is particularly interesting since the middle bands, *C2*- and *C1+*, were left unworded.

The most general Rasch analysis output, conventionally referred to as the 'facet map', 'all-facet vertical summary' or 'vertical ruler' (Linacre 2008) is given in Figure 1 below. It compares estimates of rater severity, person ability, difficulty of discourse type, and rating criteria difficulty on one scale. That is, the facet map shows rater harshness in terms of the probability of the rater awarding a given score to a test taker at a given ability. Similarly, the map displays the ability of candidates in terms of the probability of their being awarded a given score, considering what is known about the severity of the rater and the difficulty of the discourse type and rating criteria. The more able candidates are placed at the top end of the 'ruler' whereas the less able are positioned at the bottom, i.e. candidates 22 and 32 are the most able and candidates 6 and 8 are the least able in this analysis. Moreover, the map displays rating category difficulty in terms of the probability of a candidate of a given ability receiving a given score from a rater of a given severity. The most difficult category appears towards the top, i.e. *relevance, development and organisation of ideas* in this case. The five columns on the right display difficulty estimates of all scale steps in each rating criterion. For example, candidate 15 is likely to be assigned *C2*- for *lexico-grammatical resources, fluency* and *delivery* and *C2* for *relevance, development and organisation of ideas*. The measure for all probability estimates is 'logit' (log odds unit).[12]

[12] Figure 1 is a visual representation of the relative harshness of raters, the relative abilities of candidates, and the relative difficulty of discourse type as well as rating criteria. There are ten columns in this figure: one for the scale of measurement used, and one for each of the facets *raters*, *candidates*, *discourse type*, and *rating criteria*. Within the rating criteria facet, difficulty estimates of all scale levels are given in more detail on the right side of the facet map: alphanumeric strings representing level steps are positioned at integer expected scores; dashed lines (---) are positioned at expected half-score points. Raters are identified by alphabetic strings, candidates by their ID numbers. The all-facet vertical summary acts as a 'ruler' that enables us to locate and compare the facet estimates.

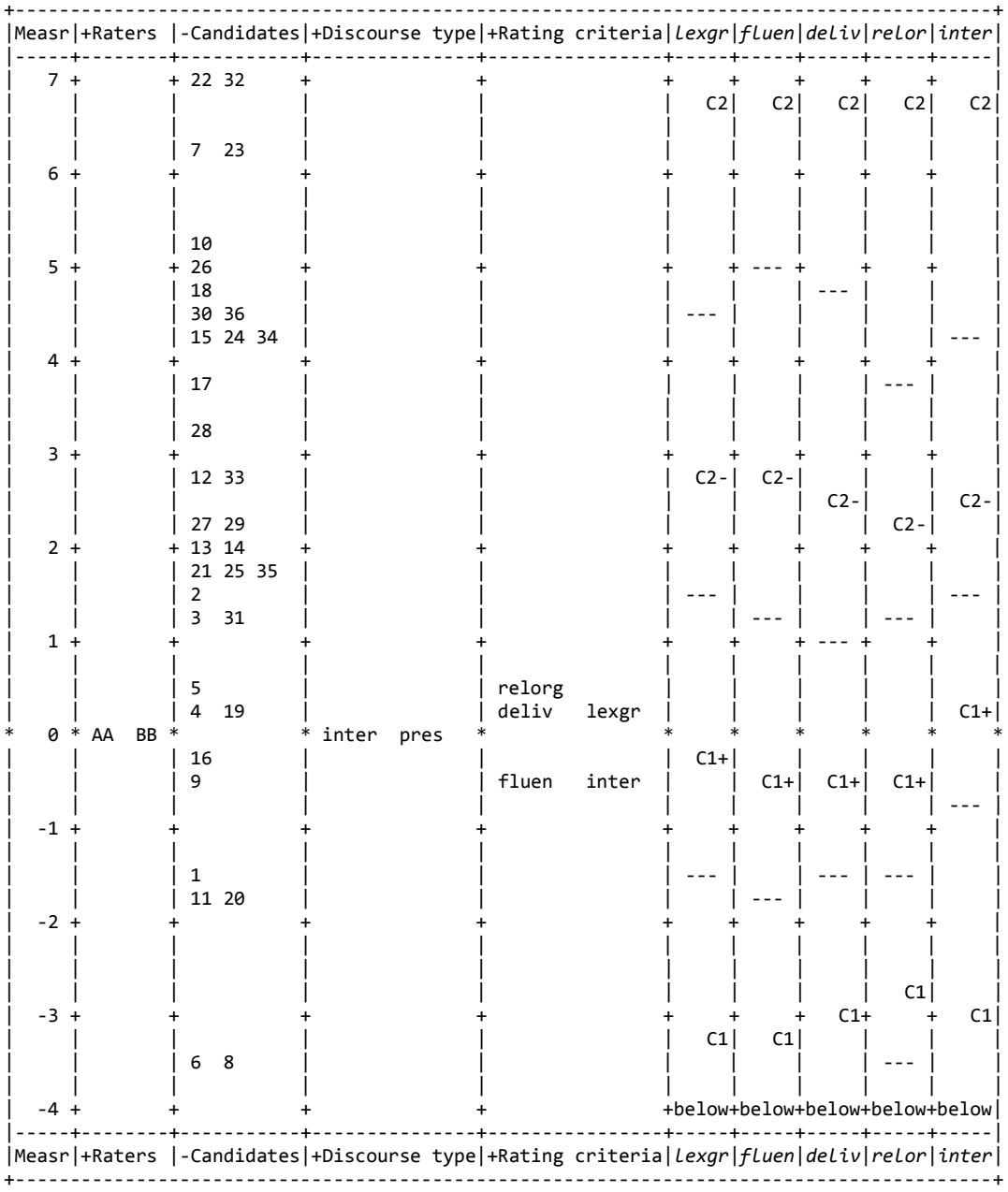


Figure 1: All facet vertical rulers

8.1 The raters

Table 1 below summarizes the results of the rater analysis. The calibrations of each rater show that the two differ only slightly in harshness by 0.22 logits

(rater AA at -0.11 and rater BB at 0.11). The numbers of error[13] associated with this measure are small (0.11 for both raters) and the fit values are good (0.98 and 1.01 for rater AA and rater BB, respectively). In traditional terms, the fit measures indicate intra-rater reliability. The expected values are 1. Where the values are closer to 0, the data are said to 'overfit' the Rasch model, i.e. there is too little variation and they are too predictable. Where the values are higher than 1, the data are said to 'underfit', which means they show unmodelled excess variation and are too unpredictable.[14] The rater reliability was also good. The reliability value reported here is not traditional inter-rater reliability, but indicates to what extent the raters act independently, as an aspect of inter-rater reliability. Near 0.0 values are preferred, since low independence signifies broad agreement. The rater reliability figure in this analysis is 0.09, which is close enough to 0.0, suggesting that the two raters are not reliably different and have similar levels of severity. The FACETS rater measurement report also shows that identical ratings were given in 155 (58.1%) out of possible 267 agreement opportunities (expected: 50.7%).

Rater	Measure (severity)	S.E.	Infit MnSq
AA	-.11	.11	.98
BB	.11	.11	1.01
Mean (Count: 2)	.00	.11	1.00
S.D.	.11	.00	.02
		Separation .32	Reliability .09

Table 1: Raters measurement report

8.2 The candidates

The second facet analysed are the candidates. Their ability measures range from -3.40 logits (candidates 6 and 8) to 8.18 logits (candidates 22 and 32). Table 2 displays the five examinees that had misfit, however, with relatively high degrees of error. As pointed out in the previous section, fit statistics help to find patterns for individual performances which do not correspond with the overall pattern. Since the number of misfitting examinees is relatively high, it

[13] Since the estimates of rater severity, candidate ability and item difficulty are extrapolations from the data available, they are subject to error. Accordingly, estimates of the likely error are provided for each measure. Ideally, the size of error should be small.

[14] There are no hard-and-fast rules for the interpretation of fit statistics. According to McNamara (1996:173), values in the range of 0.75 to 1.3 are generally acceptable. Linacre (2008:191) suggests that a range from 0.5-1.5 is productive for measurement.

is advisable to go back to the video-taped responses of the candidates and find an explanation for the ones causing the disturbance. McNamara (1996) points out that explanations should be considered in terms of failure of mastery of a particular area (diagnostic feedback), failure of attention in the test-taking situation, anxiety and the like. In general, misfitting responses suggest that the individual's abilities are not being measured appropriately by this particular test instrument.[15]

Candidate	Measure (ability)	S.E.	Infit MnSq
14	1.98	.41	.39
3	1.31	.41	1.96
4	.31	.41	.14
19	.31	.41	2.59
16	-.37	.41	1.70
Mean (Count: 36)	2.34	.53	.99
S.D.	2.84	.33	.45
		Separation .5.44	Reliability .97[16]

Table 2: Candidates measurement report (misfitting candidates)

8.3 The scale and the rating criteria

The final section of the analysis reports the results of the rating criteria measurement. Table 3 shows that *interaction* was the easiest category with a measure of -.57 logits, whereas *relevance, development and organisation of ideas* as well as *delivery* were the most difficult categories with a difficulty level of .51 and .37 logits, respectively. Standard error is relatively small for all rating categories. Furthermore, Table 3 shows that the fit values are also acceptable.

15 A re-examination of the video-taped performances is recommended, but goes beyond the scope of this analysis.

16 The separation value is 5.44, which means that the candidates can be separated into five levels. The person reliability is .97, which is the Rasch equivalent to the KR-20 Alpha statistic.

Rating criteria	Measure (difficulty)	S.E.	Infit MnSq
Relevance, development, organisation of ideas	.51	.20	1.49
Delivery	.37	.15	.83
Lexico-grammatical resources	.13	.15	.88
Fluency	-.44	.16	.82
Interaction	-.57	.24	1.42
Mean (Count: 5)	.00	.18	1.09
S.D.	.43	.03	.30

Table 3: Rating criteria measurement report

The analysis of the category statistics provides information on how well the levels of the rating criteria were distinguished. As expected, average measures increase in size as the category level increases. The average measures increase monotonically, indicating that on average the level steps are associated with a progression of candidate ability. In other words, higher abilities are awarded higher category levels. Like the average measures, step calibrations should increase gradually, too. As this pattern is not violated, and the distances between adjacent thresholds increase by at least 1.4 but no more than 5 logits,[17] it can be said that each step of the scale defines a distinct position of the variable and that the categories function well. Figure 2 shows the probability curves of the well-functioning five-level scale for *lexico-grammatical resources* as an example.[18] This figure is a visual form of investigating the distinction between level thresholds. Each category should have a distinct peak in the probability curve graph, indicating that each category is indeed the most probable score for some portion of the measures variable. The clear peaks and the separation between the categories indicate the clarity with which the scale was applied.

17 Recommendation by Linacre, quoted in Bond and Fox (2007: 224).

18 This probability curve depicts the likelihood for any score to be chosen (along the vertical axis) at any ability level (along the horizontal axis). For example, a candidate with an ability of approximately 1 logit has a probability of about 60% of receiving score 3 (*C1+*), about 25% of getting score 4 (*C1*) and about 5% of getting score 2 (*C2-*) or 5 (*below C1*).

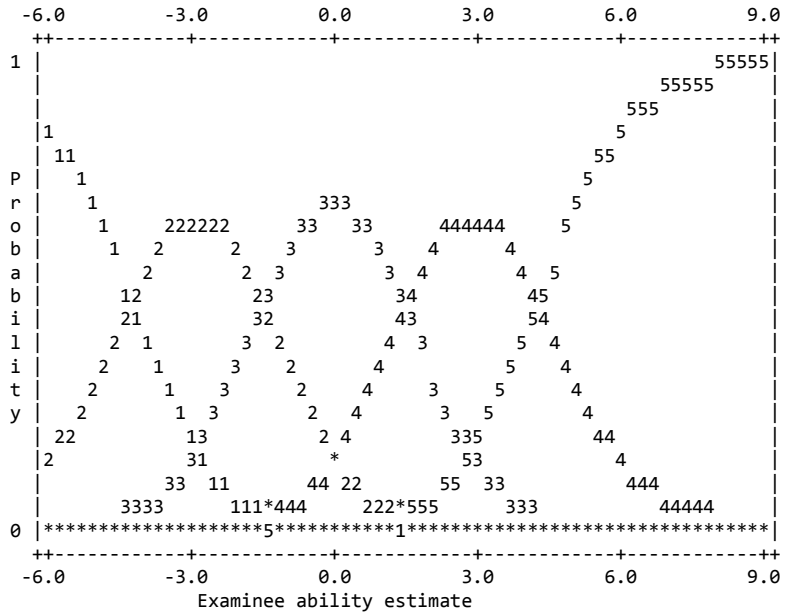


Figure 2: Probability curves of the scale for *lexico-grammatical resources*

9. Discussion

Although the scope of this study was somewhat limited in terms of sample size, and more observations, particularly ratings from different raters, would be needed to produce estimates of greater stability, the results of the multi-facet Rasch analysis have shed some light on how the new PPOCS 2 rating scale functions. The research questions posed earlier in this report addressed (a) the effectiveness of the analytic rating scale and (b) the degree of model fit of the facets. From the results it seems that most level steps in the five categories function well. Generally, all level steps are associated with a progression of candidate ability, and the raters used all level bands for the test population. Band five (*below C1*) was used as well, albeit infrequently, which was expected since most candidates taking PPOCS 2 can be assumed to have achieved a proficiency level beyond C1 according to the CEFR.

Category statistics show good step calibrations for most rating criteria. The scale for *relevance, development and organisation of ideas*, however, might need further investigation, because the steps of this scale are not consistent with those of other scales. A horizontal comparison of the scales presented in the all facet summary in Figure 1 shows that the distance between threshold estimates of band 2 and the neighbouring band 3 is rather small. The step calibrations increase only by 1.18 logits, which is considerably shorter than the distances between other threshold estimates and not in accordance with the recommended guideline of a minimum increase of 1.4 logits. In other words, the descriptor of band *C1* for *relevance,*

development and organisation of ideas is more demanding than descriptors of other categories. It might be worth considering why this is the case. For example, the two raters might have interpreted this category as more demanding than necessary. One possible remedy would be to find explicit descriptors for the in-between category steps *C1+* and *C2-* for greater clarity. However, these results should be interpreted cautiously since there is a standard error of the step calibration of .4 and more observations would be needed for more stable estimates.

While the difficulty level of the *relevance, development and organisation of ideas* criterion may warrant further investigation, the fit statistics of each facet suggest no problematic misfit for raters and the five rating criteria. Too many candidates, however, show person misfit. This may indicate that their abilities are not being measured appropriately. Closer inspection of these candidates' responses might be needed to find explanations for the disturbance.

In summary, it can be said that each analytic criterion generally seems to function well, defining distinct points on the variable and making meaningful steps in progression. Although the central level bands *C2-* and *C1+* were left unworded, it seems that raters were able to distinguish meaningfully between them. This is an indication of the possibility that the CEFR levels C1 and C2 for speaking can be further divided into more subtle yet distinguishable levels. Research will have to show whether greater explicitness and exact wordings of the in-between levels would change the results significantly.

10. Conclusion

The present study aimed to describe the four-stage development process of the new PPOCS 2 analytic rating scales and investigate the data obtained from the first live administration in February 2009, thereby gathering validity evidence for the scales. Firstly, rating criteria were intuitively selected and drafted into two analytic five-point rating scales, one for formal presentations and one for interaction. Secondly, the draft descriptors were refined according to the feedback obtained from a number of informants. After trial runs with a representative student sample in the third stage, the scales were used live and the data obtained from this administration was analysed quantitatively. On the condition that validity of a scoring procedure is defined fairly narrowly as showing good model fit, the multi-faceted Rasch analysis has offered some validity evidence. The analysis demonstrated appropriate reliability of the rating scale, which can be seen as a suitable instrument to test speaking in this context.

What follows from Messick's approach, however, is that validity is "an evolving property and the validation is a continuing process" (Messick 1989: 13) so that the development of this rating instrument should not end at this point. Considering scale validation as an ongoing endeavour, it becomes clear that this study can only be a small part of the scale validation process. It must be acknowledged that many aspects of validity, including theory-based or consequential features, are under-represented in this study and that this falls short of Messick's unified concept of validity. Consequently, the results should be interpreted cautiously in the light of their limitations. Even if the findings indicate that the inferences from the scales about speaking proficiency are valid enough and that the scales can be used as a framework for language assessment, further investigation is needed.

In fact, one of the major cautions that need to be understood refers to the limited scope of this study. The validity evidence presented here is based on only one major source of information – PPOCS 2 teachers and their interpretations of the scales. That is, other potentially relevant informants such as students were ignored. It goes without saying that the number of observations included in this analysis was limited and so the study should be replicated, involving more candidates and, in particular, more ratings from different raters. As a matter of fact, the video-recorded performances should be rated by all other PPOCS lecturers and these ratings should be included in the analysis. A larger number of raters would clearly produce estimates of greater stability. Besides, it would be intriguing to see how raters not directly involved in the design process interpret the rating scale categories. The fact that the two raters here were native-speakers of English also raises the question of whether ratings by non-native speakers of English would yield different results.

Not only the limited number of observations included in this study but also methodological constraints warrant further research. Although the scale development process included intuitive, qualitative and quantitative elements, more systematic triangulation of methods would be desirable to produce stronger results. Weir and Roberts (1994) advocate such triangulation of methods where possible. For instance, qualitative methods such as concurrent verbal protocols that focus on the raters' perception of the descriptors while rating performances or retrospective feedback questionnaires may provide further insights into the validity of the assessment process. Even more importantly, however, the scale needs an empirical underpinning to show that the descriptors of each rating category really match the candidates' performance on a given level. This approach requires a discourse-based analysis of performance and the description of key features of that

performance. In this way any significant mismatch between the level band descriptors and representative samples of performances can be discovered. The advantage of this approach is that the present rating scale can be revised and refined by very concrete descriptions based on data.

Much as further research is required and desired, this study is a major step towards more professional testing at Austrian language departments. Rarely before has the development of a testing instrument at the Department of English and American Studies at the University of Vienna received this much attention, let alone psychometric analysis. It is hoped that this study has produced some valuable results that serve as a good starting point for further scale modifications. As explained above, the development of a testing instrument is an ongoing process that is cyclical and iterative in nature and requires continuous re-assessment. The findings presented here help to see what improvements need to be made to the rating scale or the administrative processes surrounding it.

In terms of language teaching, it is hoped that this study leads to a better understanding of what is involved in speaking, teaching and testing a foreign language at tertiary level, which in turn might result in a reconsideration of current instructional practices. Indeed, a better understanding of the nature of oral language ability and a clear idea of the construct to be measured can help teachers guide and redirect their teaching, bringing essential communication skills into sharper focus and enabling to give more specific feedback on the learning progress. While students are awarded potentially fairer ratings for their performances, teachers can have more confidence in their testing practice and arrive at more informed judgements about their students' abilities. Overall, teachers and students alike might find language testing, which is an inherent part of learning and teaching, more rewarding, since greater accountability for decisions about individuals based on test results leads to more accuracy and fairness. Ultimately, it is hoped that this study can more deeply involve classroom testing practice in issues of language testing and thus contribute to professionalism in this field.

References

Alderson, Charles. 1991. “Bands and scores”. In Alderson, Charles; North, Brian (eds.). *Language testing in the 1990s: the communicative legacy*. London: Macmillan, 71-94.

Bachman, Lyle. 1990. *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, Lyle; Lynch, Brian K.; Mason, Maureen. 1995. “Investigating variability in tasks and rater judgements in a performance test of foreign language speaking”. *Language Testing* 12, 238-252.

Bachman, Lyle; Palmer, Adrian. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: OUP.

Baker, Rosemary. 1997. *Classical test theory and item response theory in test analysis*. LTU Special Report No 2. Lancaster: Centre for Research in Language Education, Lancaster University.

Berry, Vivien. 2007. *Personality differences and oral test performance*. Frankfurt: Peter Lang.

Bond, Trevor G.; Fox, Christine M. 2007. *Applying the Rasch model: fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.

Brindley, Geoff. 1991. “Defining language ability: the criteria for criteria”. In Anivan, Sarinee (ed.). *Current developments in language testing*. Singapore: Regional Language Centre, 139-164.

Brown, Annie. 2003. “Interviewer variation and the co-construction of speaking proficiency”. *Language Testing* 20, 1-25.

Brown, Annie. 2005. *Interviewer variability in oral proficiency interviews*. Frankfurt: Peter Lang.

Butler, Frances; Stevens, Robin. 1998. *Initial steps in the validation of the second language proficiency descriptors for public high schools, colleges, and universities in California: writing*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

CEFR = Council of Europe. 2001. *Common European framework of reference for languages: learning, teaching, assessment.* Cambridge: CUP. Available online at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Cronbach, Lee J. 1971. “Test validation”. In Thorndike, Robert L. (ed.). *Educational measurement*. Washington, D.C.: American Council on Education, 443-507.

Davies, Alan. 1990. *Principles of language testing*. Oxford: Blackwell.

Davies, Alan; Brown, Annie; Elder, Cathie; Hill, Kathryn; Lumley, Tom; McNamara, Tim. 1999. *Dictionary of language testing*. Cambridge: CUP.

Fulcher, Glenn. 1993. “The construction and validation of rating scales for oral tests in English as a foreign language”. PhD thesis, Department of Linguistics and English Language, Lancaster University.

Fulcher, Glenn. 1997. “The testing of L2 speaking”. In Clapham, Caroline; Corson, David (eds.). *Language testing and assessment* (Vol. 7). Dordrecht: Kluwer Academic Publishers, 75-85.

Fulcher, Glenn. 2003. *Testing second language speaking*. London: Pearson Longman.

Hatch, Evelyn; Lazaraton, Anne. 1997. *The research manual: design and statistics for applied linguistics*. Boston: Heinle and Heinle.

Henning, Grant. 1987. *A guide to language testing: development, evaluation, research*. New York: Newbury House.

Hughes, Rebecca. 2002. *Teaching and researching speaking*. London: Longman.

Kelley, Truman L. 1927. *Interpretation of educational measurement*. New York: Macmillan.

Kunnan, Anthony J. 1995. *Test taker characteristics and test performance: a structural modelling approach*. Cambridge: CUP.

Lado, Robert. 1961. *Language testing*. London: Longman.

Linacre, Mike. 2008. *A user's guide to FACETS: Rasch model computer programs*. Chicago, IL: MESA Press.

LTC = Language Testing Centre/Sprachtestcentrum, Universität Klagenfurt. 2009. "About the LTC". Available online at http://www.uni-klu.ac.at/ltc/inhalt/145.htm. (10 June 2009).

Matthews, Margaret. 1990. "The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations". *English Language Teaching Journal* 44, 117-121.

McKay, Penny. 2000. "On ESL standards for school-age learners". *Language Testing* 17, 185-214.

McNamara, Tim. 1996. *Measuring second language performance*. London: Longman.

McNamara, Tim; Lumley, Tom. 1997. "The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings". *Language Testing* 14, 140-156.

Messick, Samuel. 1989. "Validity". In Linn, Robert (ed.). *Educational measurement* (3rd ed.). New York: Macmillan, 13-103.

Messick, Samuel. 1995. "Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning". *American Psychologist* 50, 741-749.

Messick, Samuel. 1996. "Validity and washback in language testing". *Language Testing* 13, 241-256.

Milanovic, Michael; Saville, Nick; Pollitt, Alastair; Cook, Annette. 1996. "Developing ratings scales for CASE: theoretical concerns and analyses". In Cumming, Alister; Berwick, Richard (eds.). *Validation in language testing*. Clevedon: Multilingual Matters, 15-38.

North, Brian. 1995. "The development of a common framework scale of descriptors of language proficiency based on a theory of measurement". *System* 23, 445-465.

North, Brian. 2000. *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, Brian. 2003. "Relating assessments, examinations, and courses to the CEF". In Morrow, Keith (ed.). *Insights from the Common European Framework*. Oxford: OUP, 77-90.

North, Brian; Schneider, Günther. 1998. "Scaling descriptors for language proficiency scales". *Language Testing* 10, 217-262.

O'Sullivan, Barry. 2000. "Exploring gender and oral proficiency interview performance". *System* 28, 373-386.

O'Sullivan, Barry. 2006. *Modelling performance in oral language tests: language testing and evaluation*. Frankfurt: Peter Lang.

O'Sullivan, Barry. 2008. "Notes on assessing speaking". *Cornell University Language Research Centre*. http://www.lrc.cornell.edu/events/past/2008-2009/papers08/osull1.pdf (15 May 2009).

Pollitt, Alastair; Murray, Neil L. 1996. "What raters really pay attention to". In Milanovic, Michael; Saville, Nick (eds.). *Performance testing, cognition and assessment*. Cambridge: CUP, 74-91.

Shohamy, Elana. 1988. "A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10, 165-179.

Shohamy, Elana. 1994. "The validity of direct versus semi-direct oral tests". *Language Testing* 11, 99-123.

Shohamy, Elana. 1995. "Performance assessment in language testing". *Annual Review of Applied Linguistics*, 188-211.

Stansfield, Charles; Kenyon, Dorry Mann. 1992. "Examining the validity of a scale used in a performance assessment from many angles using the many-faceted Rasch model". *Education Resources Information Center*:
http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/23/f1/7d.pdf (20 April 2009).

Tyndall, Belle; Kenyon, Dorry Mann. 1996. "Validation of a new holistic rating scale using Rasch multi-faceted analysis". In Cumming, Alister; Berwick, Richard (eds.). *Validation in language testing*. Clevedon: Multilingual Matters, 39-57.

Weir, Cyril J.; Roberts, Jon. 1994. *Evaluation in ELT*. Oxford: Blackwell.

Wigglesworth, Gillian. 1993. "Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction". *Language Testing* 10, 305-336.

Is that *a filler? On complementizer use in spoken object clauses*[1]

*Gunther Kaltenböck, Vienna**

1. Introduction

This paper investigates the use of the *that*-complementizer in object noun clauses following complement-taking predicates (CTPs for short) such as *I think*. Typical examples are given in (1).

(1) a. *I think* (that) John is in Paris
b. *I suppose* (that) John is in Paris
c. *I believe* (that) John is in Paris

Epistemic phrases as these are interesting from a syntactic point of view, as they may also occur in non-initial position as syntactically detached parenthetical clauses (cf. e.g. *John is in Paris, I think* and *John, I think, is in Paris*). While in medial and final position they are clearly identifiable as juxtaposed supplements (cf. Huddleston & Pullum 2002, Peterson 1999), in initial position their syntactic status is far from clear. Are they main clauses, which syntactically govern a complement clause, or comment clauses in a parenthetical, i.e. supplementary, relationship to the following clause, which would then no longer be subordinate but a main clause? Various views have been expressed in the literature. Initial CTPs have been analysed as matrix clauses (e.g. Peterson 1999: 236, Stenström 1995: e.g. 293, 296, Svensson 1976: 375), parentheticals (e.g. Kärkkäinen 2003, Kruisinga 1932: 486, Ross 1973, Thompson 2002, Thompson & Mulac 1991), or ambiguous in status allowing both analyses depending on context and type of 'matrix' predicate (e.g. Aijmer 1972: 46, Biber et al. 1999: 197, Huddleston & Pullum 2002: 896, Quirk et al. 1985: 1113, Urmson 1952: 481).

1 I would like to thank the VIEWS team for their constructive feedback and Lotte Sommerer for pointing out useful references.

*The author's e-mails for correspondence: gunther.kaltenboeck@univie.ac.at.

The syntactic status of initial epistemic clauses is particularly difficult to determine for cases where the *that*, as an explicit marker of subordination, has been omitted. However, even when the complementizer is present the syntactic structure is far from clear as much hinges on the actual function of the *that*. This paper tries to shed some light on the syntactic status of such initial CTPs in a corpus of spoken language. As the title already indicates, the question I want to answer is an unusual, perhaps even provocative one: Could it be that in spoken language *that* no longer functions as a marker of subordination but merely as a filler, i.e. operating on the linear plane rather than indicating hierarchical difference? I would like to suggest that this is indeed the case.

To answer the question I will proceed in the following way. After a brief stock-taking of corpus frequencies of different lexical predicates in Section 2, I will first discuss different arguments for identifying the main clause status of a CTP-phrase in Section 3: Section 3.1 reviews some syntactic arguments and shows that they are inconclusive. Section 3.2 discusses cognitive-functional arguments, which suggest that, although CTPs are generally backgrounded, their status is essentially indeterminate and depends on contextual realisation. Section 4 then takes a closer look at prosodic prominence as a formal signal for foregrounding the CTP and investigates whether prosodic prominence may be indicative of a main clause status of the CTP. The prosodic analysis focuses on *I think* as a representative and extreme case of grammaticalisation and shows that there is little difference between *I think* + zero and *I think* + *that*. This parallelism together with the lack of prosodic highlighting for *I think* in both constructional types casts additional doubt on the subordinator status of *that*. This view is corroborated by further analysis of cases of *that* insertion, which suggests that it is mainly used as a filler for rhythmical purposes or to alleviate production difficulties. Section 5 offers a brief conclusion.

2. Frequency and predicate types

The data analysed for the present study are derived from the spoken section of ICE-GB, the British component of the *International Corpus of English* (Nelson et al. 2002), which comprises roughly 600,000 words. The corpus yields a number of lexical predicates that can occur in initial as well as non-

initial (i.e. parenthetical) position (cf. Kaltenböck 2006b, 2008, 2009b for details). Table 1 lists the seven most frequent predicates in the corpus.[2]

	- that	*+ that*	Total
I think	91.0% (1,036)	9.0% (102)	100% (1,138)
I suppose	93.6% (88)	6.4% (6)	100% (94)
I hope	84.6% (44)	15.4% (8)	100% (52)
I believe	47.8% (22)	52.2% (24)	100% (46)
I guess	95.0% (19)	5.0% (1)	100% (20)
I'm afraid	88.2% (15)	11.8% (2)	100% (17)
I suspect	50.0% (5)	50.0% (5)	100% (10)

Table 1: Complement-taking predicates in the spoken section of ICE-GB with and without *that*-complementizer (raw figures in brackets)

The overview in Table 1 shows that most predicates have a clear preference for contact clauses (i.e. zero *that*). With an average share of 10.7 percent the use of the complementizer represents a highly marked option.[3]

As observed in previous studies (e.g. Elsness 1984, Biber 1999, Kaltenböck 2006a) the proportion of *that* is to some extent text type dependent with more formal texts showing a higher percentage of *that* insertion. This is most noticeable with *I think*, which has only 5.1 percent of *that* insertions in Private dialogue and increasing proportions in the more formal text types Public dialogue (10.7%), Public monologue (12.1%), Scripted speech (18.6%). With the remaining predicates the results are less conclusive, presumably owing to the low number of occurrences, but even here Private dialogue shows consistently the lowest numbers (cf. Appendix for details).

If we compare the different lexical predicates, it is possible to identify two groups: those with an average percentage of less than 10 percent of *that*-insertions, viz. *I think*, *I suppose*, *I guess*, and those with higher percentages, viz. *I hope, I believe, I'm afraid, I suspect*. This distinction may be related to a semantic difference between the two groups: the former are weak assertives (Hooper 1975), whose semantic content is reduced essentially to just epistemic meaning, the latter are strong assertives, which add some more

[2] In an attempt to increase semantic and syntactic homogeneity of the group and consequently comparability of the individual predicates I have excluded those containing negation (*I don't think, I don't know*), those expressing certainty (*I'm sure, I know*), and those followed by an extraposed complement (*it seems, it appears*).

[3] Similar results have been found e.g. by Thompson & Mulac (1991) for *think* in spoken American English (91%), and Tagliamonte & Smith (2005) for *I think* in British dialects (91%).

specific semantic content, e.g. positive expectations in the case of *hope*.[4] *I believe* and *I'm afraid*, however, require some further explanation.

Clause initial *I believe* may show different degrees of assertion depending on the type of proposition it introduces. Compare the examples in (2).

(2) a. *I believe* that John is in Paris
b. *I believe* that there is a God

All other things being equal (esp. context and prosody), sentence (2a) is more likely to receive a weak assertive (i.e. parenthetical, hedging) interpretation than sentence (2b). The non-verifiable nature of the proposition in (2b) will generally favour a strong assertive (i.e. matrix clause) reading, viz. 'I assert the belief that there is a God' (rather than: 'There may be a God') (cf. Hooper 1975: 100-101, Quirk et al. 1985: 1113). Arguably, this tendency towards a strong assertive reading is reduced by the omission of *that*.

With clause-initial *I'm afraid* the distinction between weak and strong assertive reading is even more determined by the presence or absence of *that*. Compare, for instance, (3a), which will normally receive a parenthetical/hedging interpretation, viz. 'I regret to say', while (3b) is more prone to being interpreted as matrix clause with the meaning of 'I fear'.[5]

(3) a. *I'm afraid* John will be late
b. *I'm afraid* that John will be late

The general pattern of distribution in the corpus thus seems to suggest some correlation between semantic content of the predicate and explicit marking of subordination by *that*: predicates with a more definite meaning, i.e. one that goes beyond the mere expression of tentativeness, are more likely to be followed by *that* than more fully grammaticalised hedges such as *I think, I suppose, I guess*.[6] For ambiguous predicates such as *I believe* and *I'm afraid*

---

[4] In the case of *I suspect* the higher frequency of *that* may be attributed to more frequent occurrence in the formal text type Public dialogue, although traces of historically earlier uses, e.g. 'to expect with apprehension' (OED s.v. suspect v., 5) or a tinge of 'imagining something undesirable/being suspicious' (ibid., 1) cannot be excluded.

[5] Of course the type of proposition also plays a part with (i) triggering more strongly a matrix clause reading:

(i) *I'm afraid* that John will lose his job

Crucially, however, it is the tense of the 'object' clause that shapes the reading. Past tense, as in (ii), is less compatible with the meaning 'I fear' for pragmatic reasons and therefore makes such a reading less likely.

(ii) *I'm afraid* that John lost his job

[6] Diessel and Tomasello's (1999) study of the use of the that-complementizer in early child language seems to point in a similar direction. They *found that the complementizer is generally absent with evidential*

the presence or absence of *that* may be a signal for suggesting either one or the other interpretation. In other words, the corpus results in Table 1 can be taken as an indication that *that* may be used to mark the difference between a matrix- or a comment clause reading of the CTP, i.e. acting as a genuine subordinator. However, the low figures for most of the predicates do not warrant far-reaching conclusions. Moreover, this view fails to explain why an almost fully routinised and formulaic hedge such as *I think* still takes a fair amount of *that*-complementizers. In what follows I will therefore focus mainly on *I think*, not only because of the larger number of data available for analysis, but also, because as an 'extreme case' of a semantically weakened predicate, it represents a particular challenge for an explanation of the use of the complementizer.

3. Main- or comment clause?

Before taking a closer look at the use of *I think* in the corpus let me review some of the arguments put forward in the literature for the analysis of CTPs as main- or comment clauses Section 3.1 presents the main syntactic arguments. Section 3.2 considers the cognitive-functional arguments.

3.1 Syntactic arguments

Various syntactic tests have been proposed to show that there is a difference in status between CTPs taking a *that*-complementizer and those taking zero, the implication being that the former are proper main clauses whereas the latter are not. Put differently, the use of the *that*-complementizer distinguishes between main- and comment clause. I will briefly discuss three such tests and show that the evidence they allegedly present is far from conclusive.

One way of showing the difference between *that* and zero *that* clauses is by way of the **tag-question test** (e.g. Aijmer 1972: 52, 1997: 8, Hand 1993: 501, Knowles 1980: 405). The argument is that the 'subordinate' clause in (4) has lost some of its subordinate status since it allows various 'main clause phenomena' (Hooper-Thompson 1973, Green 1976), such as the tag question.

(4) *I think* ∅ John is in Paris, isn't he

markers such as I think, I guess, I bet, I mean, I know. Only three verbs are commonly used with a complementizer: *say, tell* and *pretend*. As explanation Diessel and Tomasello (1999: 96) point out that "[t]hese three verbs have a more concrete meaning than all other verbs in our sample".

However, the same seems to be true for clauses with *that*, as in (5) (cf. Aijmer 1997: 8), although Hand (1993: 501) marks its acceptability as questionable.

(5) *I think* that John is in Paris, isn't he

Conversely, the 'matrix clause' in a sentence without *that*, as in (6a) does not seem to allow questioning in this way (cf. Aijmer 1997: 8, Knowles 1980: 405). This is equivalent to the behaviour of 'real' parentheticals, as in (6b) and can be taken as an indication that the clause lacks illocutionary force.

(6) a. *I think* ∅ John is in Paris, *don't I
b. John is in Paris, *I think*, *don't I

The validity of this test, however, is questionable, as the unacceptability of the tag in (6) could also be attributed to a pragmatic restriction, viz. the inappropriateness of a speaker questioning (doubting) his/her own statement. Indeed, if we substitute a pragmatically more likely tag, as in (7), the result is acceptable in both cases.

(7) a. *I think* John is in Paris, don't you
b. John is in Paris, *I think*, don't you

Another form of **question test** is intended to show that questioning of the CTP is ok with an explicit complementizer, as in (8a), but somewhat questionable without a complementizer, as in (8b) (cf. Huddleston & Pullum 2002: 896, Asher 2000: 33).

(8) a. A: *I believe* that John is in Paris B: Really. Do you
b. A: *I believe* John is in Paris B: ?Really. Do you

Similar results can be obtained from the **negation test** (adapted from Erteschik-Shir & Lappin 1979: 46):

(9) a. *I believe* that John is in Paris
– No, that's a lie. I don't actually.
– No, that's a lie. He isn't actually.

b. *I believe* John is in Paris
– No, that's a lie. ?I don't, actually.
– No, that's a lie. He isn't, actually.

The results of these tests, however, depend on the type of predicate used. Unlike *believe*, which is well-known for its semantic ambivalence (cf. Section

2), more fully grammaticalised[7] predicates such as *I think* yield different results. Moreover, acceptability depends to a large extent on context and prosodic delivery. The above tests therefore do not provide conclusive evidence for a different syntactic status of CTP phrases with and without *that*.

A fairly strict division between main- and comment clause uses of *I think* is also drawn by Thompson and Mulac (1991), but based on corpus data rather than syntactic tests. They suggest that certain combinations of main clause subjects and verbs (such as *I think*) "are being reanalyzed as unitary epistemic phrases. As this happens, the distinction between 'main' and 'complement' clause is being eroded … with the omission of *that* a strong concomitant" (*op. cit.*: 249). This view of a syntactic reanalysis of *I think* has been criticised by Kearns (2007), who proposes a pragmatic explanation in terms of informational prominence for the presence or absence of an overt subordinator, but maintains the traditional matrix clause analysis of *I think*. Thompson (2002) herself, in a more recent publication, has moved in the opposite direction, suggesting that all CTP-phrases in conversation are best analysed as epistemic formulaic fragments rather than superordinate matrix clauses irrespective of presence or absence of a *that*-complementizer. This view is largely based on functional evidence and will be discussed in more detail in the following section.

3.2 Cognitive-functional arguments

Functional perspectives generally support a parenthetical or comment clause analysis of *I think* (and similar CTP-phrases) irrespective of whether they are followed by a complementizer or not. Kärkkäinen (2003: 41, 2009 forthc.), for instance, notes that epistemic phrases (such as *I think*) with *that* are functionally equivalent to those without. Another functional investigation is Thompson (2002), who challenges the standard analysis of CTP-phrases as matrix clauses and argues that finite indicative complement clauses (with and without complementizer) are generally not subordinate but override the 'main clause' (CTP). The CTP should therefore not be analysed as superordinate matrix clause but as epistemic/evidential/evaluative fragment, often of a formulaic nature, which simply expresses "speaker stance towards the assessments, claims, counterclaims, and proposals" (*op. cit.*: 134). In her argumentation Thompson refers to Langacker's (1991: 436ff) definition of a

[7] I am using the term grammaticalised here, as defined for instance by Hopper & Traugott (2003) or Brinton & Traugott (2005), although the term pragmaticalised, suggested by Erman & Kotsinas (1993), might be more appropriate for CTP-phrases owing to their pragmatic function.

subordinate clause as "one whose profile is overridden by that of the main clause … *I know **she left*** designates the process of knowing, not of leaving", where "profile" refers to the "relative prominence accorded to various substructures" (Langacker 1991: 4). Thompson (2002: 131) interprets the notions of "profile" and "relative prominence" in terms of the interactional actions that an utterance is performing in a particular context (cf. Goodwin & Goodwin 1992, Linell 1998, Pomerantz & Fehr 1997, Schegloff 1990). From the analysis of her corpus examples she concludes that "the talk doing the actions that the participants are jointly engaged in doing is either in a main clause turn or in a finite indicative complement" (*op. cit.*: 134), while "the CTP-phrases do not constitute the speakers' interactional agenda" (*ibid.*). The "action" or "interactional agenda" is thus roughly equivalent with the "'issue' around which the talk centers" (*op. cit.*: 133) or, presumably, the discourse topic. Example (10) illustrates her point with the CTP *I don't care* and an *if*-clause complement, which in her view is functionally equivalent to a *that*-clause (= Thompson's ex. 13; boldface indicates the talk accomplishing the action).

(10) [Frank and his young son Brett have noticed that Brett's sister Melissa appears to be about to mark on Brett's art project]

1	MELISSA:	are you gonna add like the little lines that jut out of [these]?
2	FRANK:	[get your pen] back from that
3	BRETT:	…yeah
4	MELISSA:	**it's erasable**,
5		**and I am not marking on it**.
6	BRETT:	…I don't care **if it's erasable**.
7		**don't touch it**.
8	MELISSA:	**(HI I didn't HI)**
9	BRETT:	…**I know**
10		…**don't**

Although Thompson's analysis is a compelling one, her identification of discourse prominence does not provide a satisfactory explanation for all data. A case in point is example (10) above, where prominence could also be analysed in terms of information structure. This would yield an entirely different result for the construction in question: the complement clause, which is entirely retrievable from the preceding co-text, has to be seen as informationally backgrounded, while the CTP represents the communicatively salient, i.e. new (irretrievable) bit of information, which contributes to the further development of the communication (cf. e.g. Firbas' 1992 notion of communicative dynamism). In contrast to Thompson, it is therefore the assertion of the CTP *I don't care* that is the main point of the utterance. This

problem has also been noted by Boye and Harder (2007: 576f), who conclude that epistemic stance cannot automatically be equated with secondary discourse function. In their model they consequently distinguish between "stance-marking as an aspect of lexical meaning and stance-marking as an inherently secondary, 'parenthetical' discourse or usage function" (Boye & Harder 2007: 577) (cf. below for further discussion).

It seems, however, that information structure cannot always provide a clear-cut answer either. Compare, for instance, the following example from Thompson (2002: 132), which contradicts a simple equation of informational retrievability and non-assertion.[8]

(11) [at a birthday party, after Kevin was discovered to have lettuce on his tooth, everyone has jokingly commented on it, and Kendra has asked for a toothpick]

WENDY: …everybody's getting uh, tooth obsessed
KEN: I guess we a=re.

Here the complement *we are* represents given (retrievable) information, which is clearly reflected in its elliptical form. Nonetheless, the main point of the utterance is not the CTP *I guess* but the complement. The reason is that what is at issue here (the 'action' in Thompson's terms) is the act of agreeing or affirming the previous utterance, for which the CTP, owing to its semantic vagueness, is not a suitable candidate. The example illustrates that establishing the communicatively salient part of the construction is not always a simple straightforward matter of equating givenness with backgrounding but has to take into account a variety of factors: 'interactional action', information structure, and the semantic value of the predicate.

In spoken language, however, there is an additional means available to the speaker to signal prominence: prosody. And it is prosodic highlighting which in example (11) above will help to identify the complement as the main point of the utterance rather than the CTP *I guess*. As noted for instance by Halliday (1985: 277), new or 'newsworthy' information is information that is presented by the speaker as such. Prosodic prominence is such a means of presentation (I will return to prosodic prominence in the subsequent section).

The examples above show that from a cognitive-functional perspective the distinction between main- and comment clause may depend on a variety of factors and is far from clear-cut: instead of a neat binary distinction, a functional view suggests a gradient link between main- and comment clause.

---

8 Although Boye and Harder seem to argue for a view of prominence in terms of newness (cf. Boye & Harder 2007: 576) in contradistinction to Thompson, they, somewhat surprisingly, fully accept Thompson's analysis of this example.

Such a scale has been proposed by Boye and Harder (2007), who identify the following three categories, which are seen as three different stages in the development of CTPs such as *I think*: (1) primary lexical CTPs, (2) secondary lexical CTPs, and (3) secondary grammatical CTPs. This classification takes into account both the structural and the usage status, each of which is described by a binary set of values: lexical vs. grammatical *structural status* and primary vs. secondary *usage status*. While the first stage is easily identifiable as matrix clause and the last stage as (clause internal and final) comment clause, the second stage is a hybrid category, which exhibits a discrepancy between usage status and structural status and as such is descriptively ambiguous (cf. Boye & Harder 2007: 586). *I think* in clause-initial position seems to qualify for precisely this intermediate stage: structurally, its morphosyntactic form is that of a lexical clause (Boye & Harder 2007: 591) and its syntactic position that of a prototypical matrix clause. However, in terms of its discourse function initial *I think* is typically secondary.[9]

A similar view is presented by Nuyts (2000: 122ff), who sees epistemic modal expressions such as *I think* as a "battleground" where two conflicting functional forces are at work: an information structural force and an iconic (or conceptual semantic) force. From the perspective of iconicity the status of the epistemic evaluation is that of an operator (i.e. a meta-representational element) over a state of affairs, which suggests main clause status for the epistemic expression "since it directly reflects the meta-status of the qualification relative to the state of affairs" (Nuyts 2000: 123). In terms of information structure, on the other hand, the epistemic qualification is backgrounded and the state of affairs foregrounded, i.e. it carries the focal information. The information structural force therefore works against a main clause interpretation for the epistemic expression, since main clauses prototypically carry foregrounded information and embedded clauses backgrounded information (cf. Brandt 1984, Givón 1984, Mackenzie 1984, Sadock 1984, Tomlin 1985).

Taking up Nuyts' metaphor, we can think of clause-initial epistemic markers such as *I think* as 'undecided battles' where the different forces outbalance each other and allow for different interpretations of the status of *I think*. In other words, the result of these conflicting forces is one of

9 Cases where the epistemic qualification of *I think* is the main point of the utterance are possible but extremely rare. Cf.

(i) A: So you're telling me John is in Paris
B: I THINK John is in Paris (but I'm not sure)

neutralisation and indeterminacy in the sense of Boye and Harder's hybrid category 'stage 2'. In spoken language, as pointed out above, an additional force enters the 'battleground' and may 'tip the scales', viz. prosody. As an iconic reflection of prominence (cf. Bolinger 1985) prosodic signals may be seen as decisive factor for the interpretation of the syntactic status of the CTP. The following section therefore takes a closer look at the prosody of initial *I think*.

4. Prosodic analysis

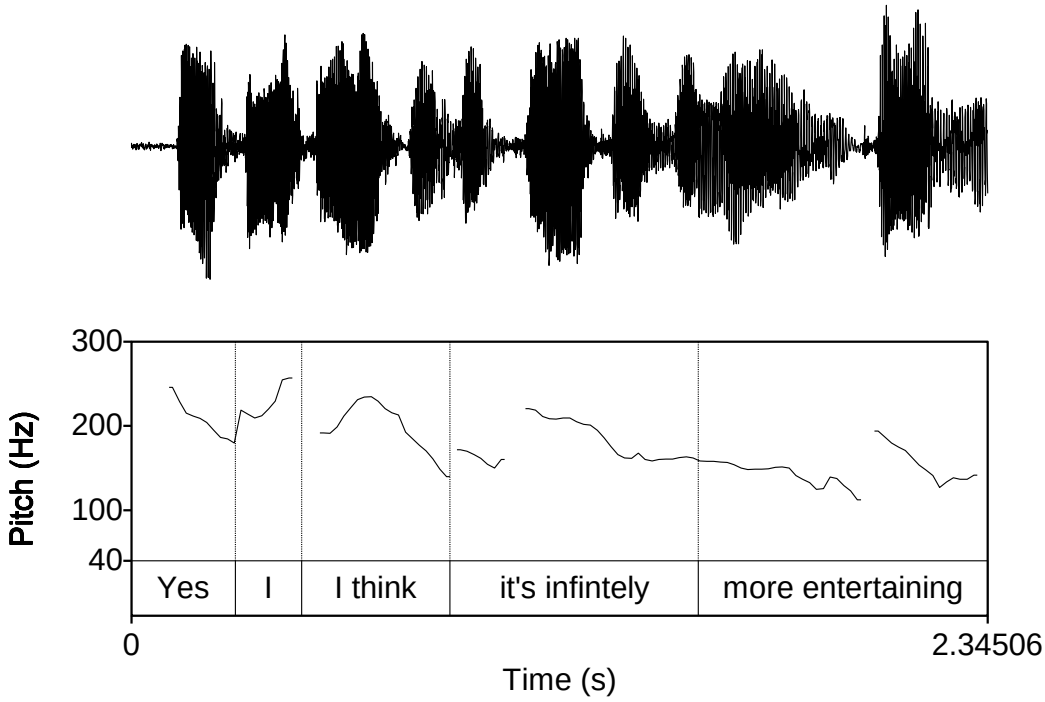
This section provides a detailed prosodic analysis of initial *I think* with a view to establishing whether its prosodic realisation suggests foregrounding and therefore a matrix clause interpretation. This is interesting, as a matrix clause analysis of *I think* would entail a classification of *that* as a subordinator. Section 4.1 first analyses the prosody of *I think* + zero, which is then compared in Section 4.2 with the results for *I think* + *that*.

4.1. *I think* + zero

In spoken language, prosody is a prime indicator of functional prominence, with prosodic prominence iconically reflecting the communicative salience of a linguistic element. Prosodic prominence is, however, not simply a matter of 'high' or 'low' but of degrees. In a previous study, which distinguishes between prosodically bound and independent comment clauses, I have identified three different degrees of prosodic prominence for comment clauses in initial position: (i) separate tone unit, (ii) part of the head, or (iii) part of the pre-head (Kaltenböck 2008: 95ff, cf. also Kaltenböck 2009a). Each of theses types is illustrated below for *I think*.

(i) *I think* with an **independent tone unit** is exemplified in (12), which has a nuclear accent on *think* and is followed by a tone unit boundary, indicated by a change in pitch level (cf. Cruttenden 1997: 35 on boundary markers). As a possible alternative the nucleus may also be on the pronoun *I* rather than on the predicate *think* (cf. Simon-Vandenbergen 2000: 50; Kaltenböck 2009 forthc. for the function of such uses).

(12) Yes *I I think* it's infinitely more entertaining (s1b-024-12)

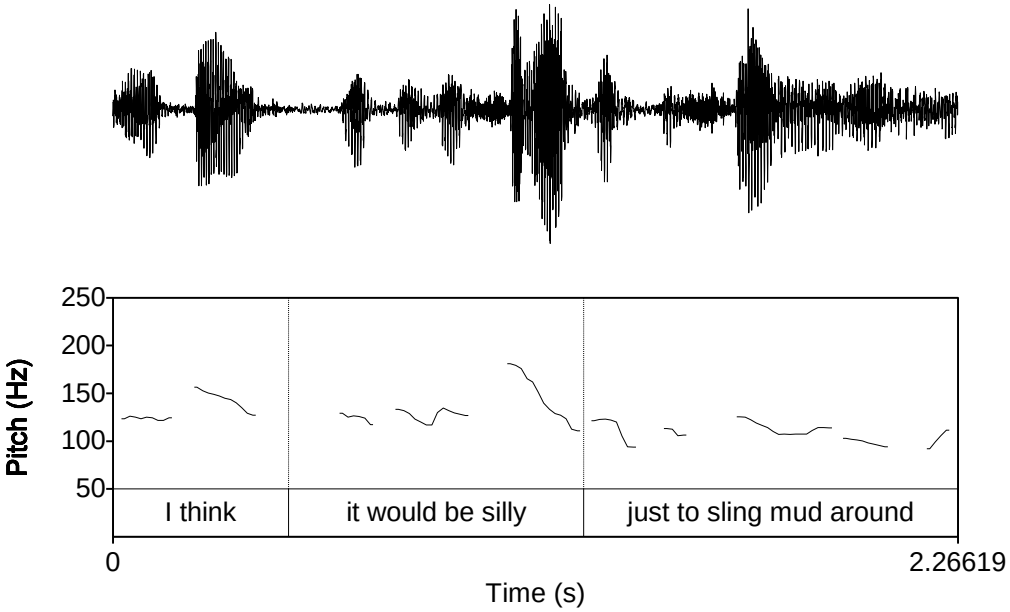


(ii) An example of *I think* **integrated into the head** is given in (13), where *think* represents the first accented syllable in the tone unit, the so-called onset (e.g. Wells 2006: 207) but is less prominent than the nuclear accent on *silly* (cf. Cruttenden 1997: 54 for a definition of head).[10]

[10] To distinguish between heads and nuclei the following criteria were applied:

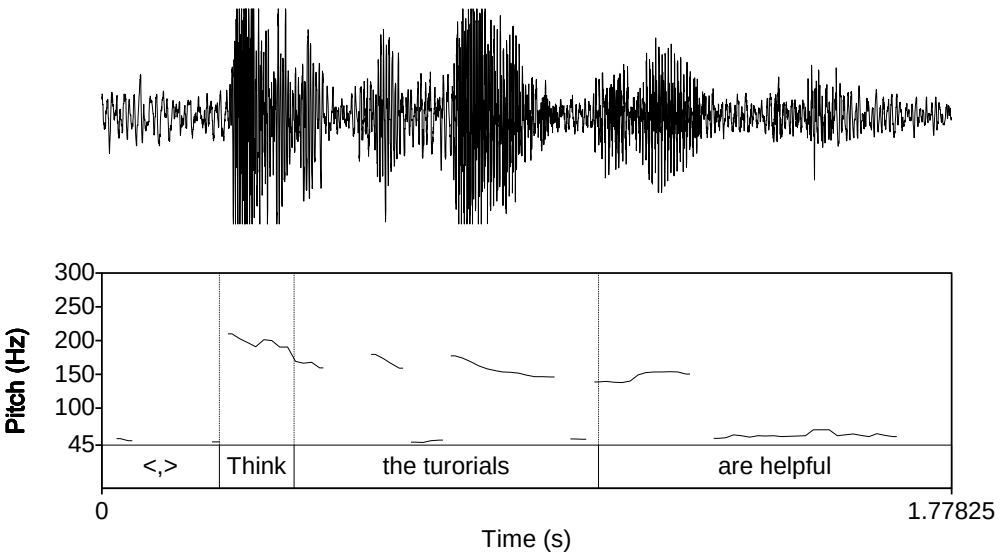
a. Onset syllables are generally on a higher pitch level than the nucleus owing to declination within a tone unit, i.e. the fact that pitch tends to be lower at the end of a tone unit than at the beginning (e.g. Couper-Kuhlen 1986: 82-83, Wichmann 2000: 103-105).
b. If at the beginning of a tone unit, i.e. not preceded by a pre-head, the onset will often be anacrustic, i.e. produced with greater speed (cf. Cruttenden 1997: 32).
c. Only in case of a separate nucleus is *I think* followed by a tone unit boundary, as indicated by features such as anacrusis, final syllable lengthening, change of pitch level or pitch direction of unaccented syllables (cf. Cruttenden 1997: 35).
d. Onsets are less prominent than nuclear accents, which is reflected phonetically in a smaller range of pitch movement and/or weaker energy pulses.

(13) *I think* it would be silly just to sling mud around (s1b-022-19)



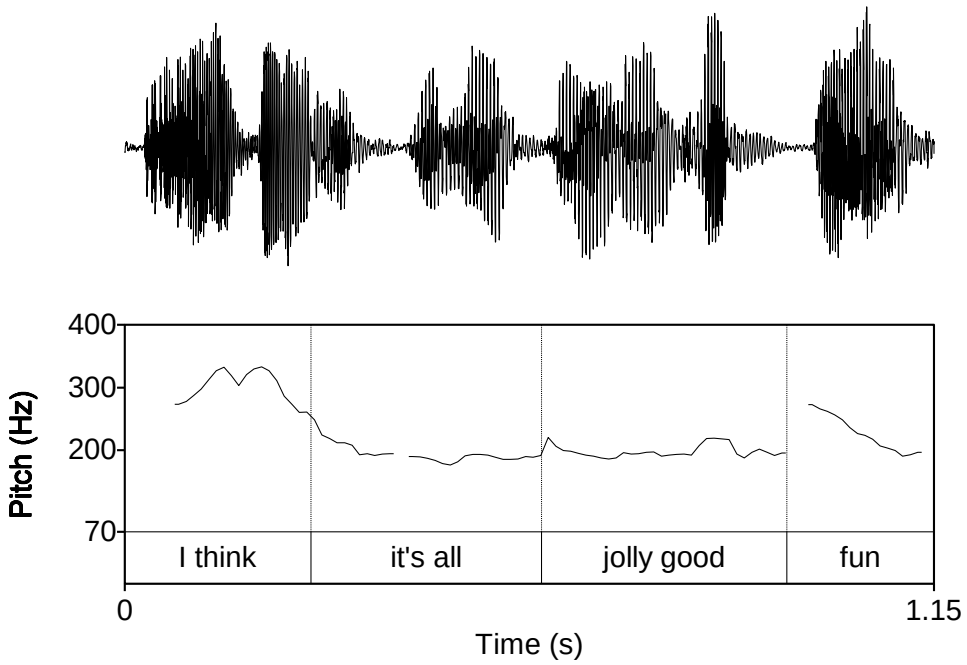
Typically in such cases the accent will be on the predicate *think* as in example (13) above. As unstressed element, *I* represents the pre-head but may be suppressed altogether as in (14) (where <,> indicates a short pause).

(14) *Think* the tutorials are helpful (s1b-015-4)



Occasionally, however, the accent occurs on the *I* (rather that on *think*), which then starts the head and gives the *I* an implicit contrastive interpretation (*I* as opposed to someone else), as in example (15).

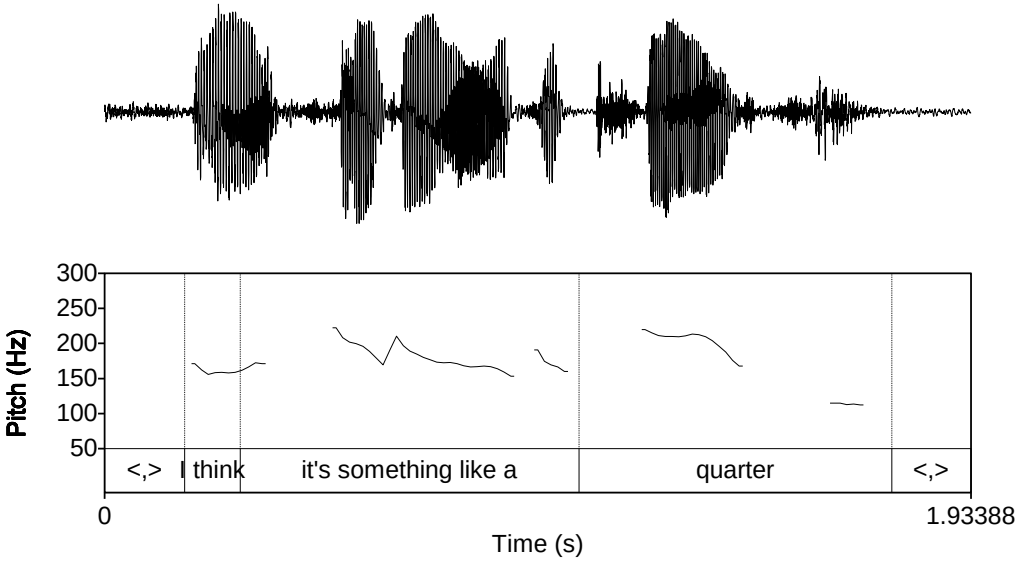
(15) *I think* it's all jolly good fun (s1b-024-28)



(iii) The third prosodic pattern is that of integration in the form of a **pre-head**, i.e. an unaccented (typically unstressed and anacrustic) syllable preceding the head (cf. Wells 2006: 214-15).[11] This pattern is exemplified in (16), where the string *I think it's* forms the pre-head, followed by an accented syllable *some*, which starts the head, and the nucleus on *quarter*.

11 The term 'stress' is used here as rhythmically stressed, while 'accent' refers to a syllable made prominent by rhythmic stress and pitch prominence, i.e. by a change in pitch, movement in pitch, or the start of a pitch movement (cf. Wells 2006: 93).

(16) *I think* it's something like a quarter (s1b-030-29)



For the corpus analysis only subsection Public dialogue (s1b) in ICE-GB was taken into account, which is the only text category that has a sufficiently large number of *that*-clauses (viz. 52, cf. Table A in the Appendix). The prosodic analysis of *I think* + zero is based on 148 random instances (of a total of 434 in Public dialogue), which were analysed both auditorily and instrumentally with the help of the acoustical analysis programme PRAAT (Boersma & Weenink 2008). The results are summarised in Table 2 below.

	n	%
Prosodically independent	7 (of which nucleus on *I*: 3)	4.7%
Right-bound: part of head	112 (of which accent on *I*: 9)	75.7%
Right-bound: part of pre-head	29	19.6%
Total	148	100%

Table 2. Prosodic patterns of initial *I think* in Public dialogue followed by a zero *that*-clause

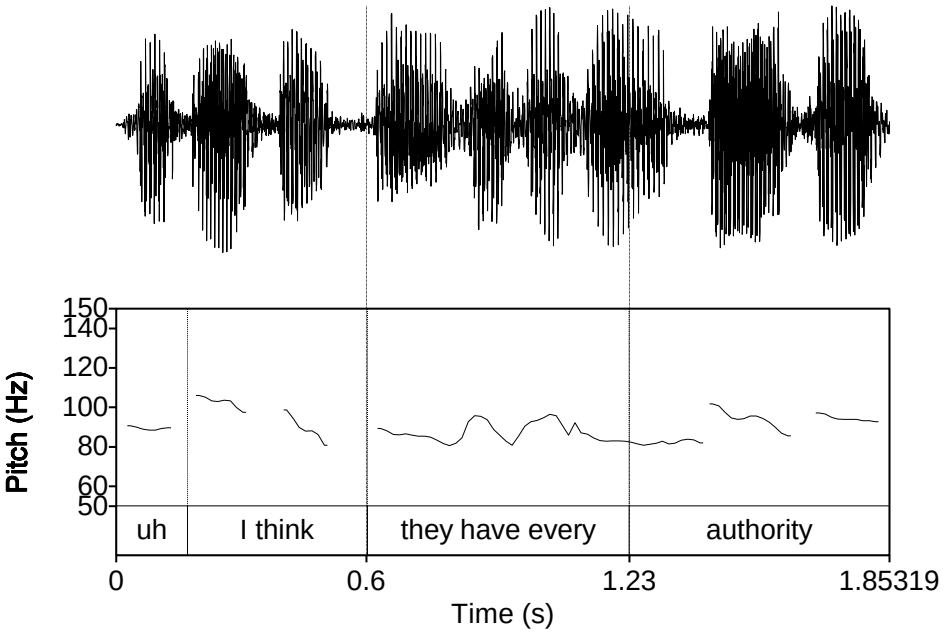
We can see that the dominating pattern is that of *I think* being realised as part of the head (75.7%), followed by its realisation as pre-head (19.6%). An independent tone unit for *I think* is extremely rare (4.7%). This lack of nuclear prominence is, however, not really surprising, as *I think* is clearly the most grammaticalised (pragmaticalised) of all comment clauses and has therefore been subject to a high degree of semantic bleaching (e.g. Mindt 2003). This semantic reduction makes *I think* an unlikely candidate for nuclear highlighting.

In previous studies I have identified various functions of comment clauses and *I think* in particular (Kaltenböck 2008, 2009 forthc.), showing that comment clauses can be further grammaticalised from epistemic markers into pleonastic structuring devices (cf. also Van Bogaert 2006). These uses tend to be phonetically reduced and lack prosodic prominence. Initial *I think* realised as pre-head can be equated with this structural function.

What about the remaining prosodic realisations for *I think*, nuclear tone and head accent? If, as noted above, we take prosodic prominence as an iconic reflection of the syntactic status of *I think*, it is tempting to correlate prosodic prominence in the form of a separate nuclear tone (tone unit) with matrix clause status and reduced prominence in the form of a head with comment clause status. Such a correlation of syntactic status with prosodic prominence would also seem to fit in with the presumed diachronic development of comment clauses, which are seen by Thompson and Mulac (1991) and Traugott (1995) to have started out as matrix clauses which have grammaticalised into epistemic markers/comment clauses (and further into discourse markers with filler function; cf. Kaltenböck 2008, 2009b). Although Thompson and Mulac's matrix clause hypothesis has been dismissed by Brinton (1996) and Fischer (2007a, b), who suggest a derivation from adverbial clauses (cf. *as I think*), it may still be assumed that the starting point was a fully lexical item, i.e. Boye and Harder's (2007) primary lexical CTPs (cf. Aijmer's 1997 full lexical meaning 'cogitation').

However, while a simple correlation of prosodic prominence with the syntactic status of *I think* may be intuitively appealing and may have indeed some theoretical value, it falls short of providing a complete explanation for the corpus data. Simply correlating hierarchical status, i.e. main- vs. comment clause, with degrees of prosodic prominence ignores the fact that prosody not only has 'vertical' function in the sense of foregrounding/backgrounding or *mise en relief*, but may also have linear or 'horizontal' function by linking and rhythmically structuring elements of speech. A closer look at the corpus data shows that there are indeed cases where prosodic prominence seems to have been prompted by rhythmic considerations. Compare, for instance, example (17), where the separate chunking of *uh I think* as an independent tone unit with nucleus on *I* may have been triggered by an implicit desire to conform to a rhythmical pattern which involves chunks of roughly 6 milliseconds: */ uh I think / they have every / authority /*

(17) Uh *I think* they have every authority both from their governments and from the UN resolutions to do that (s1b-027-103)

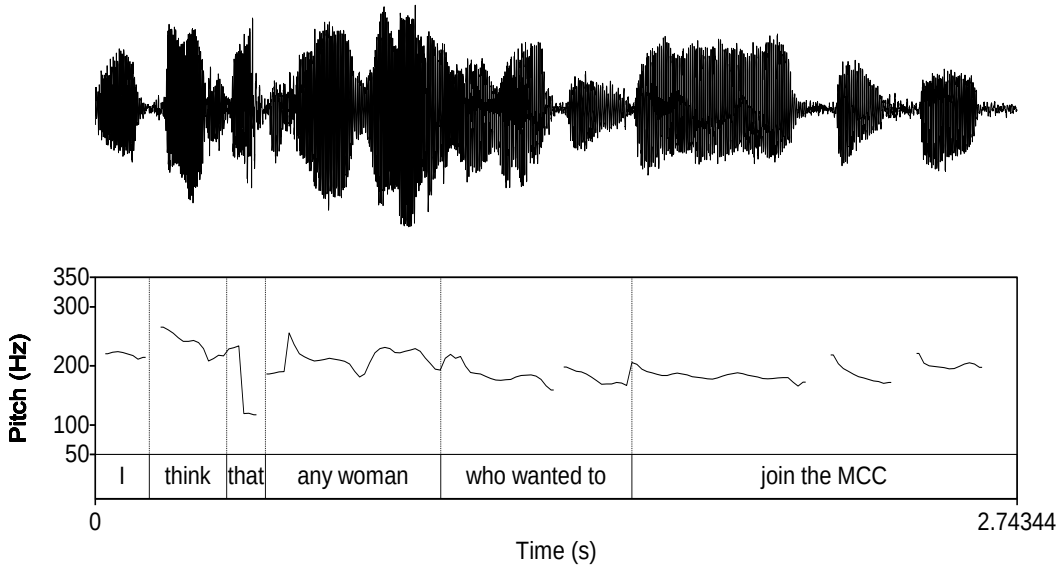


This temporal or linear aspect may be more prominent where *I think* is used in a hesitation phase as a staller, whose function is to 'buy time'. Giving *I think* more prominence (e.g. a nuclear rather than an onset accent) may allow the speaker to do precisely that. It may also be assumed that rhythmic considerations come more into play in public speaking with experienced speakers (i.e. the text category under investigation). Note also that the insertion of a *that*-complementizer in the above example would disrupt the regularity of the rhythm. I will discuss this issue in more detail in the following section.

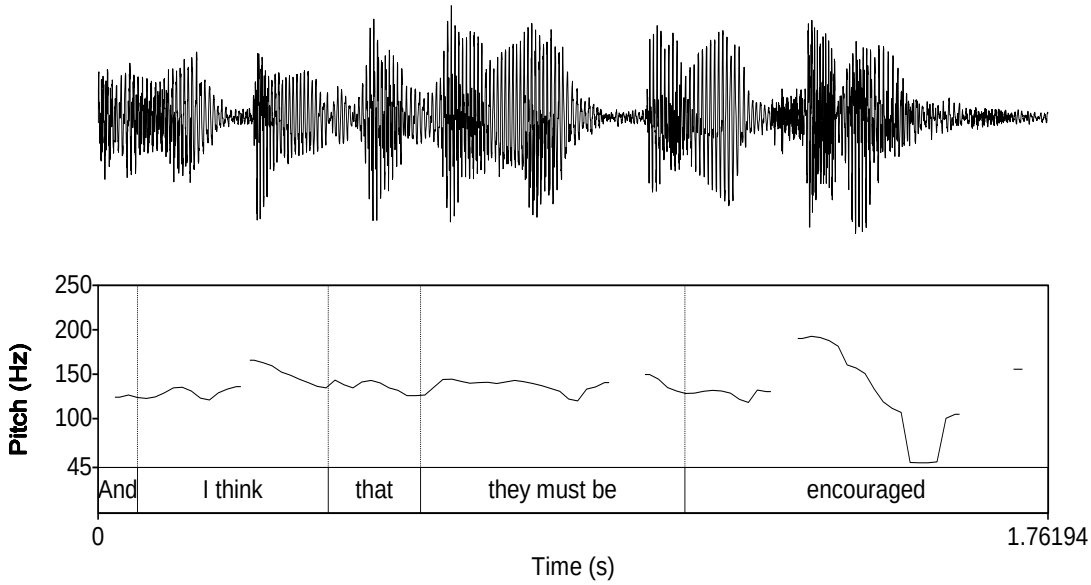
4.2. *I think + that*

In contrast to *I think* + zero, *I think* followed by the subordinator *that* could be taken as indication for main clause status of *I think*. The prosodic analysis, however, suggests otherwise. The analysis of all 52 instances of *I think* + *that* in Public dialogue (s1b) shows, first of all, that the same three patterns can be identified as for the zero clauses above, viz. (i) nuclear accent, (ii) accented syllable in the head, and (iii) pre-head. These three patterns are illustrated by the examples in (18), (19) and (20) respectively.

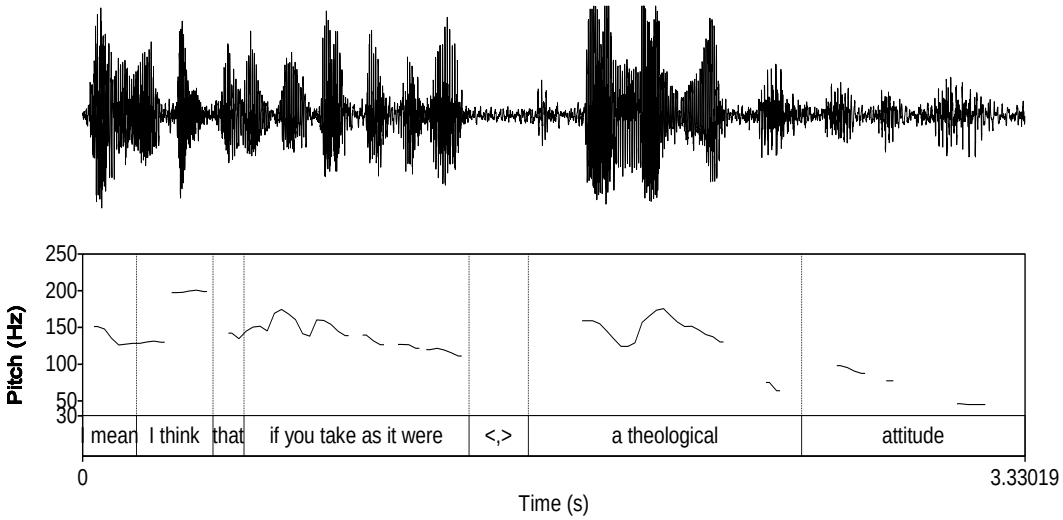
(18) *I think* that / any woman who wanted to join the MCC (s1b-021-26)



(19) And *I think* that they must be encouraged (s1b-036-72)



(20) I mean *I think* that if you take as it were a theological attitude (s1b-039-93)



In example (18) *think* takes a **nuclear tone** with a tone unit boundary after the complementizer, as indicated by the pitch change on *any*. In example (19), on the other hand, *think* represents the **onset of the head**, which leads up to (and includes) the initial syllable of *encouraged*. *Think* is preceded by the unstressed syllables *and + I*, which represent the pre-head. In example (20) the **pre-head** includes both *I mean* and *I think*, with the head starting on *that*. As noted for zero *that*-clauses, the accent (both nuclear and non-nuclear) may shift away from *think* to the pronoun *I*, as for instance in example (21) below.

If we compare, as a next step, the distribution of the three prosodic patterns for *I think* + *that*-clause with that of *I think* + zero, we find that they closely correspond. Table 3 shows that the most frequent pattern by far is again that of heads (75%), followed by pre-heads (13.5%) and independent tone units (11.5%).

	n	%
Prosodically independent	6 (of which nucleus on *I*: 2)	11.5%
Right-bound: part of head	39 (of which accent on *I*: 5)	75.0%
Right-bound: part of pre-head	7	13.5%
Total	52	100%

Table 3. Prosodic patterns of initial *I think* in Public dialogue followed by a *that*-clause

If we take nuclear prominence in the form of a separate nuclear tone as a possible cue for main clause status (as discussed in the previous section), we have to conclude that the prosodic realisation of *I think* provides no indication

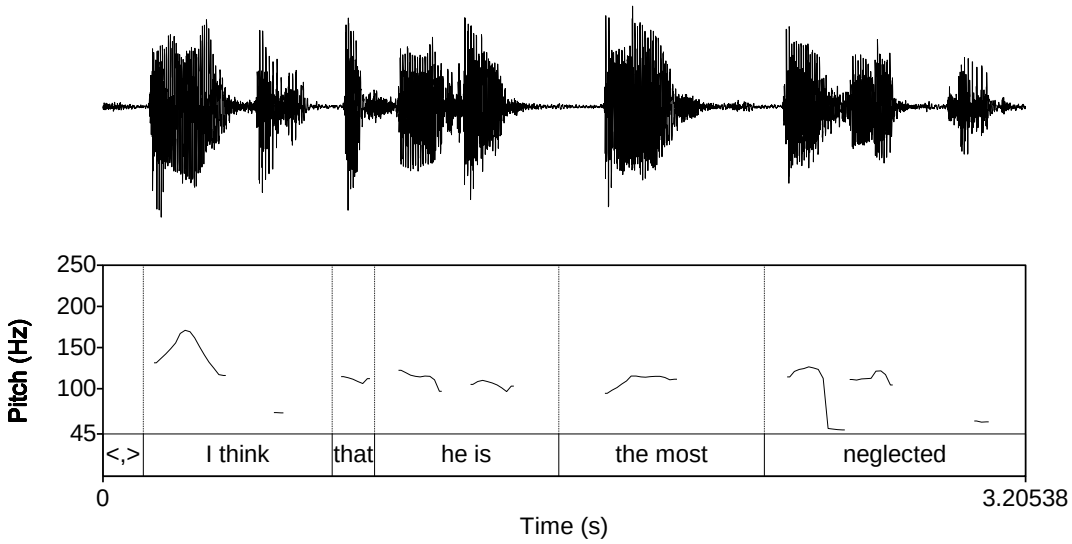
for a general main clause status of *I think* + *that*, only for a small minority of instances (11.5%). This may be somewhat surprising as *that* is a marker of subordination and therefore points at main clause status of *I think* on the structural level (cf. Boye & Harder 2007). On the usage level, however, *I think* + *that* generally has secondary status, as indicated by its preferred prosodic realisation: its reduced prosodic prominence signals that *I think* is not normally the main asserted content of the utterance. Moreover, the parallel distribution of the three prosodic types for *that* and zero clauses suggests the there is no fundamental difference in usage between the two constructional variants. This confirms an assumption that has already been variously expressed in the literature, for instance by Kärkkäinen (2003, 2009 forthc.) or Nuyts (2000: 129 note 13).[12]

On the level of usage, therefore, overt marking of subordination by a *that*-complementizer does not make a significant difference in terms of prosodic foregrounding of *I think*. The only difference is that with zero clauses we find a somewhat higher percentage of pre-heads and lower percentage of nuclear accents, which can be taken as an indication of *I think* + zero having moved even further down the path of grammaticalisation. Overall, however, the distributional pattern is similar, which, in turn, raises the question whether *that* still functions as a marker of hierarchical difference between the two clauses. I will discuss this point in more detail in the following.

Let us take a closer look at the prosodic realisation of the *that*-complementizer itself. It shows that it can be intonationally grouped either with *I think* or the following clause. This difference in association is most obvious in cases where *I think* carries its own nuclear tone and is therefore followed by a tone unit boundary. This boundary may either associate *that* with *I think*, as in example (18) above, or associate it with the following clause, as in (21).

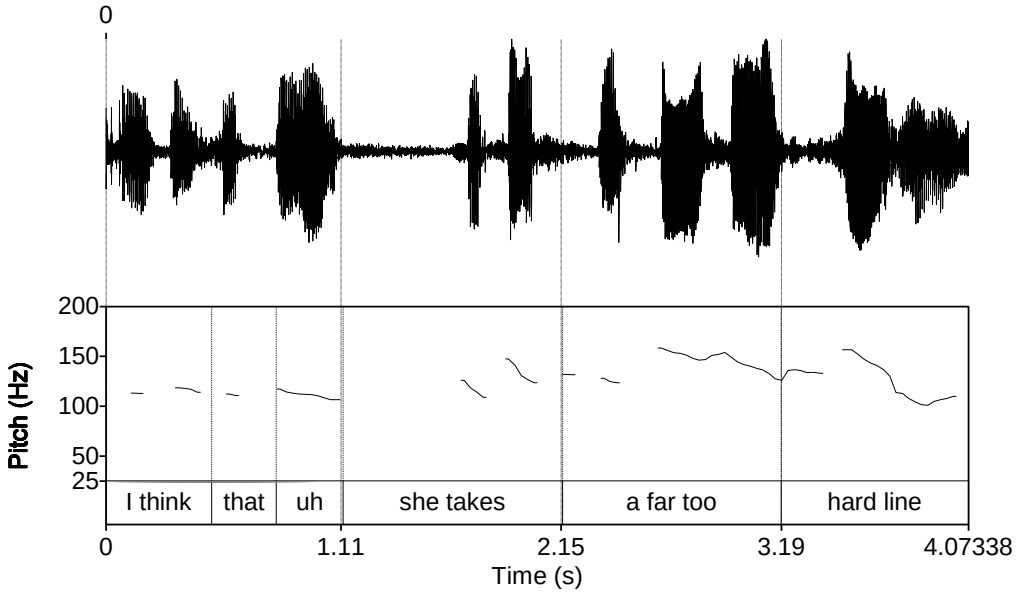
12 This view is also implied, but not overtly expressed, in Thompson (2002) and Boye and Harder (2007).

(21) *I think* / that he is the most neglected of that uh number of composers around the turn of the century (s1b-032-103)



In head and pre-head realisations of *I think* the complementizer is typically integrated into the larger pitch contour but may occasionally also show signs of association or dissociation with *I think*, albeit less markedly so. Compare for instance example (20) above where *that* is part of the head with example (22) below, where it is part of the pre-head.

(22) *I think* that uh she takes a far too hard line (s1b-035-20)



The prosodic realisation of *that* therefore does not necessarily reflect the syntactic analysis of the construction, which identifies the complementizer as part of the subordinate clause.[13] Such a mismatch between syntax and prosody is not really surprising and has been noted before for various other constructions (e.g. Brazil 1997, Wichmann 2001). It is interesting, however, that there is a tendency for *that* to be prosodically grouped with *I think* rather than the following clause.[14]

How can we explain this lack of correspondence between syntax and prosody? Associating the complementizer on the usage level with *I think* (and indeed inserting it in the first place) seems to result from the speaker wanting to add weight to the CTP in the form of an extra syllable.[15] The reason for this may be twofold:

First, adding an extra syllable to the initial clause makes it longer and therefore more effective as clause initial staller used for bridging a hesitation phase, which is one of the main functions of initial *I think* (cf. e.g. Stenström 1994, 1995). Compare, for instance, the following example where *that* has a staller function similar to that of *uhm* and *and* (cf. also example 22 above).[16]

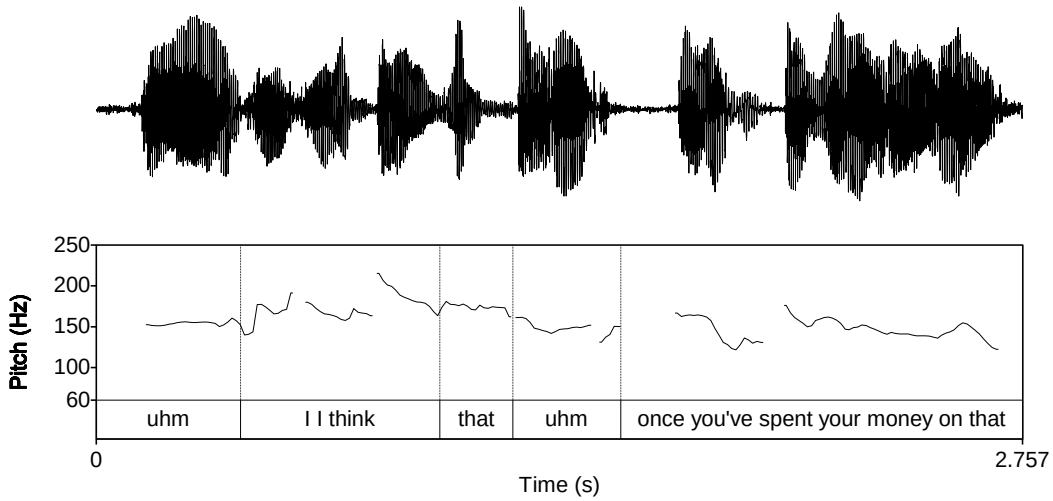
13 Cf. however the semantic analyses by Davidson (2001), Lepore & Loewer (1989) and Hand (1993) for a different view.

14 This is also reflected in the fact that in the corpus *I think* and *that* are never separated by intervening material (e.g. hesitation sound, filler), whereas *that* is frequently separated from the clause of which it is the head. The level of performance therefore seems to suggest a closer association of *that* with the main clause rather than the subordinate clause.

15 With two syllables *I think* is one of the shortest of all comment clauses (cf. Kaltenböck 2006b, 2008), which incidentally also seems to have contributed to its advanced stage of grammaticalisation.

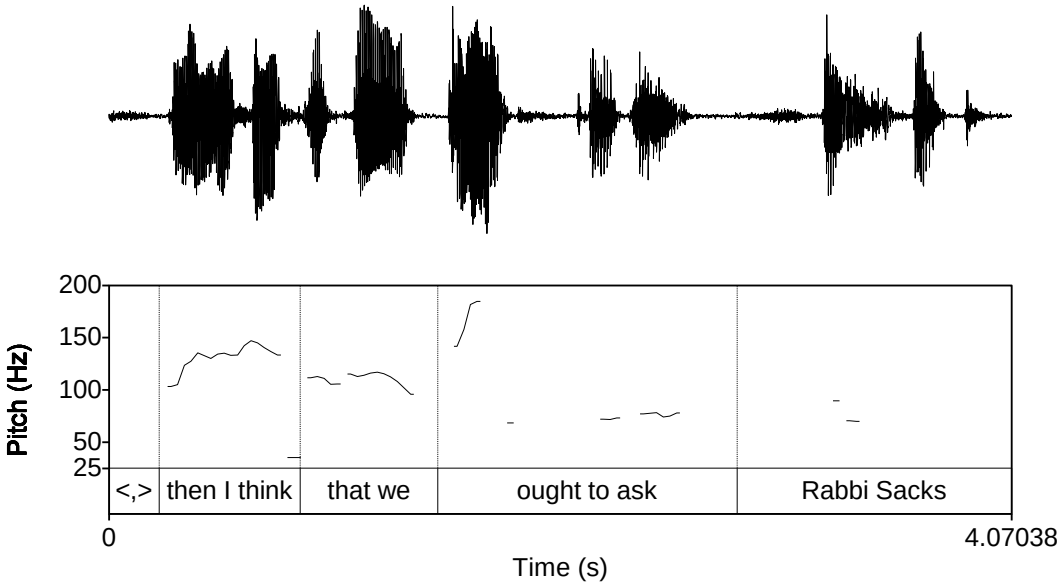
16 The nuclear accent on *think* in this example may in fact have resulted from the stalling function of *I think*: giving it more emphasis allows the speaker to gain time and extend the hesitation phase. As noted in Section 4.1, prosodic prominence not only has a hierarchical/foregrounding function but also a temporal/linear one.

(23) Uhm *I I think* that uhm once you've spent your money on that the thing to spend your money on is a subscription to the local horticultural society (s1b-025-133)



Second, prosodic association or dissociation of *that* with *I think* can be motivated by rhythmic considerations. This is illustrated by example (24), where chunking *that* with material following it rather than material preceding it results in two rhythmic chunks of roughly equal length: *then I think* and *that we*.

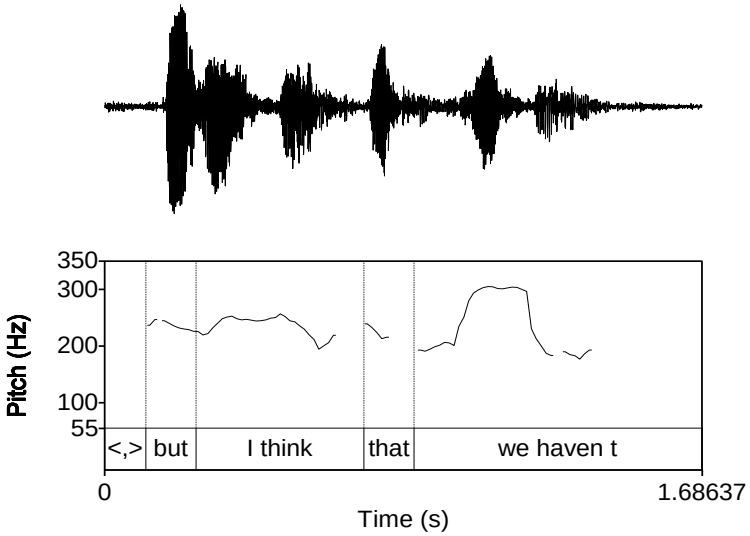
(24) then *I think* that we ought to ask Rabbi Sacks t uh uh to uh uh to say more because of course he has said two important things (s1b-028-63)



Occasionally, rhythmic chunking of *that* is underpinned by parallel intonation contours as in example (25), where association of *that* with the following

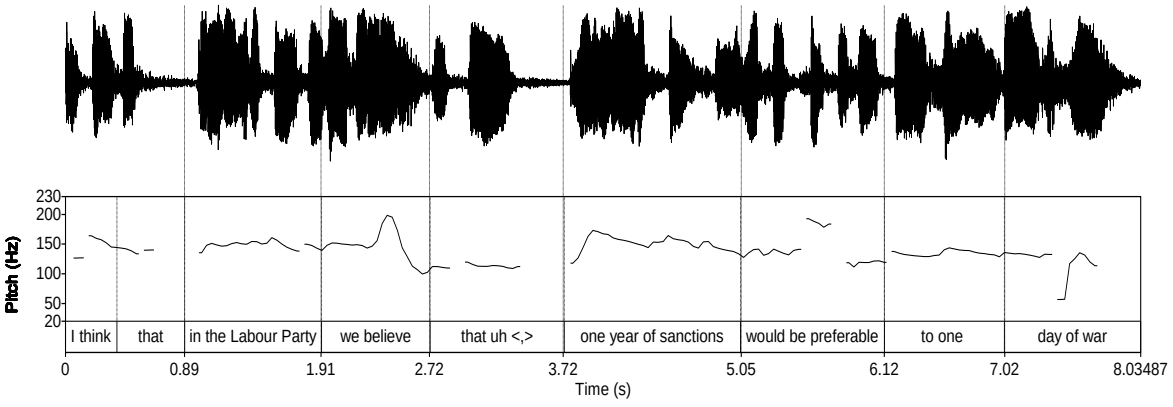
clause creates two three-syllable chunks (*but I think* and *that we haven't*), each with the same fall-rise-fall intonation contour.

(25) But *I think* that we haven't in the sense that we have just classification still (s1b-012-104)



The underlying principle for rhythmic chunking seems to be that of rhythmic harmony, viz. a tendency towards rhythmic chunks of roughly equal size (cf. principle of isochrony). This is illustrated in example (26), where association of *that* with *I think* brings the first rhythmic unit in line with the average length of the following ones, i.e. roughly 1 millisecond (note incidentally the same length of the second hesitation phase: *that* + *uh* + <,>).

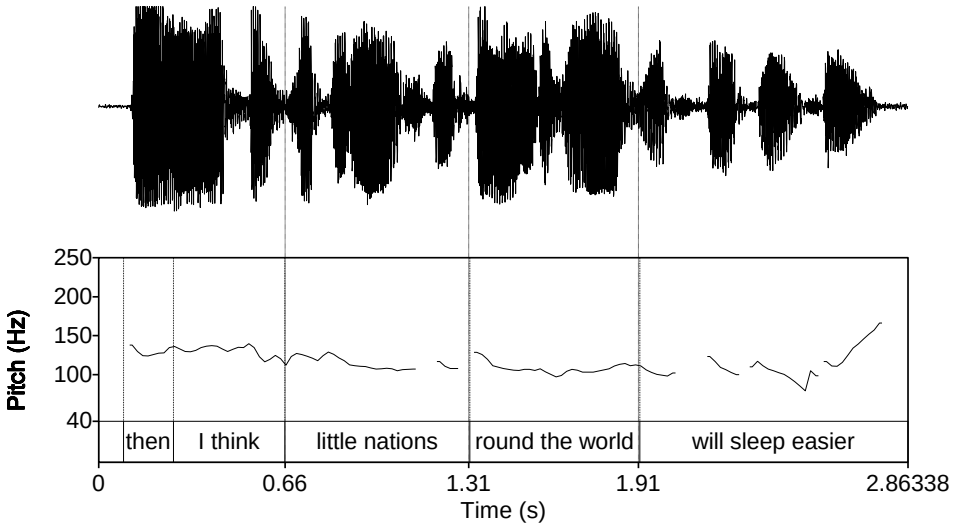
(26) *I think* that in the Labour Party we believe that uh <,> one year of sanctions would be preferable to one day of war (s1b-035-29)



While it is clear that the principle of rhythmic harmony cannot be pressed too far, it seems that the text type of public conversation is particularly susceptible to it, especially the text categories broadcast discussions and broadcast interviews, which typically involve highly experienced public speakers and incidentally have the highest proportion of *that* in the corpus (6.6 and 3.2 occurrences per 10,000 words respectively as opposed to 1.2 occurrences for Private dialogue).

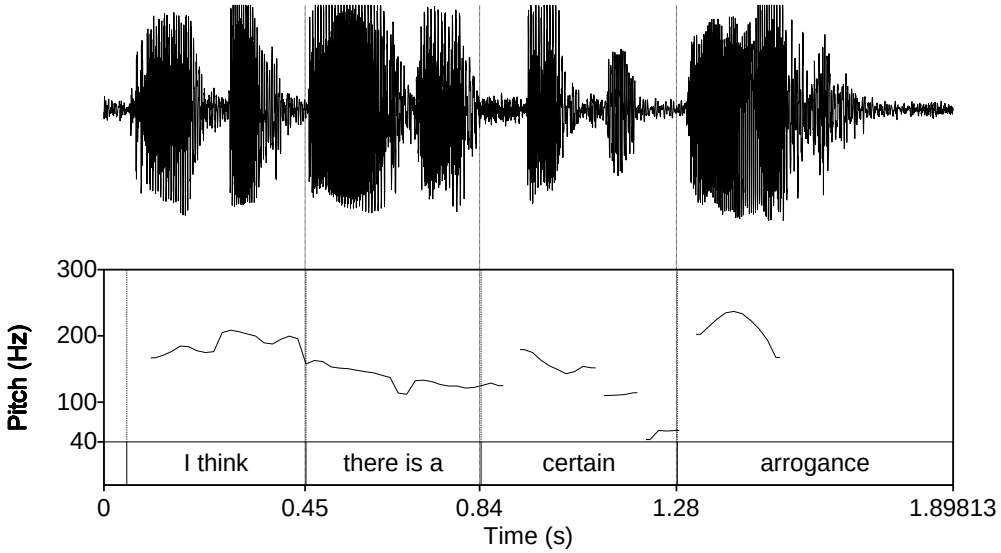
Note, however, that the function of adding extra weight to initial *I think* and making it conform to an overall rhythmical pattern is by no means restricted to *that*. Other items may fulfil a similar function, such as *then* in example (27).

(27) Then *I think* little nations round the world will sleep easier (s1b-027-69)



The rhythm of the construction can also be linked to the type of subject in the *that*-clause. Consider, for instance, example (28), where the subject of the second clause consists of an unstressed syllable (*there*), which is followed by two further unstressed syllables (*is a*). The resulting rhythmic pattern of the entire construction *I think there is a certain arrogance* (which has an accent on *I*) is thus: ● ●● ● ●●●●●●. Note that a *that*-complementizer would add an extra unstressed syllable to the row of three unstressed syllables, which is not desirable for rhythmical reasons.

(28) *I think* there is a certain arrogance on the part particularly of the extreme left in Britain on this matter (s1b-027-136)



A closer analysis of the corpus data shows that unstressed subjects in the form of existential *there* + *be* or the pronoun *it* + *be* strongly prefer omission of *that*: with these subjects *that* is used in only 3.1 percent (3/91) in Public dialogue (s1b), in contrast to all other subjects where *that* occurs 12.6 percent of the time (49/389). This preference for *that*-omission in fact proves to be significantly affected by the presence of the unstressed subjects *there* + *be* and *it* + *be*: $\chi^2 = 6.41 >$ crit (df = 1, p = 0.05).

Similar results are obtained if we take into account all pronouns that are typically unstressed, i.e. existential *there*, anticipatory *it*, and all personal pronouns, but disregard all other types of subject, such as full NPs, clauses, stressed pronouns (e.g. *this, that, mine*) as well as pronouns preceded by intervening adverbials or hesitation markers, i.e. pronouns not immediately following *I think (that)*. The statistical analysis shows that for initial *I think* in Public dialogue a preference for zero *that* is highly significantly affected by the presence of these unstressed subjects (cf. Table 4).

	- *that*	+ *that*	Total
Unstressed pronominal subjects (personal pronoun, existential *there*, anticipatory *it*)	211 (94.2%)	13 (5.8%)	224 (100%)
Other subjects (full NP, clause, stressed pronoun, pronoun preceded by adverbial	223 (85.1%)	39 (14.9%)	262 (100%)
Total	434	52	486
$\chi^2 = 9.095 >$ crit (df = 1, p = 0.01)			

Table 4. Occurrence of *that* with unstressed pronominal subjects in the 'object' clause

This finding ties in with Elsness' (1984) observation that complex subjects correlate with *that*-retention. As a possible explanation for this he notes that "[a]lthough there is no risk of ambiguity in such constructions, one may see the selection of *that* connective as a contribution to greater syntactic clarity" (Elsness 1984: 532). This may be true for written texts. For spoken language, however, it is necessary to take into account rhythmic considerations, viz. unstressed subjects favouring *that*-omission and, closely associated with this, memory constraints in online production: production of a syntactically complex subject, which can also be expected to have high informational value, will normally require extra 'thinking time' (cf. Rohdenburg's 1998 complexity principle), which is provided for by the *that*-complementizer.[17]

Elsness (1984) also mentions adverbials occurring at the boundary between matrix verb and object clause subject as a factor favouring *that* insertion. He attributes this to "a (conscious or unconscious) desire on the part of the writer to avert ambiguity" (Elsness 1984: 532). In other words, *that* insertion identifies the adverbial as belonging to either the matrix- or the object clause. In the case of spoken *I think* (in Public dialogue), however, all adverbials preceding the subject of the object clause are clearly part of the *that* clause. There are no instances where *that* insertion would indicate association of an adverbial with the CTP. *That* always immediately follows *I think*. With *that* omission, on the other hand, all adverbials in pre-subject position (adverbs, PPs, clauses) are unambiguously identifiable as part of the 'object' clause on semantic (and grammatical) grounds, cf. for instance (29).

(29) a. *I think* [according to your evidence] Ferndale Business Services got in touch with you … (s1b-064-97)

b. Uhm <,> eh uh *I think* [when I was younger] I was more self-confident and arrogant and perhaps ruthless you know ... (s1b-041-204)

Disambiguation therefore can be excluded as a conditioning factor for *that* insertion with spoken *I think*. The high proportion of *that* omission (80.9%) with pre-subject adverbials also attests to this. Nonetheless, *that* insertion is still significantly affected by adverbials preceding the 'object' clause subject (cf. Table 5). Compare, for instance, example (30).

17 Elsness (1984) notes coreferentiality of the pronominal object clause subject with the matrix clause subject as a further conditioning factor for *that* omission in written texts. For spoken *I think*, however, coreference of the two subjects does not play a major role: only 4.6 percent (20/434) of all zero *that*-clauses in Public dialogue have *I* as their subject, compared to 1.9 percent (1/52) of *I*s in *that*-clauses.

(30) And *I think* that that [perhaps in the lectures] there was there wasn't really a a a an appreciation of the positive benefit to religious traditions of the cultural engagements which took place (s1b-028-37)

	- *that*	+ *that*	Total
Adverbials preceding 'object' clause subject	55 (80.9%)	13 (19.1%)	68 (100%)
No preceding adverbials	379 (90.7%)	39 (9.3%)	418 (100%)
Total	434	52	486
$\chi^2 = 5.13 >$ crit (df = 1, p = 0.05)			

Table 5. Occurrence of *that* with adverbials preceding the 'object' clause subject

The reason why pre-subject adverbials favour *that* insertion again seems to lie in the greater syntactic complexity of the 'object' clause. Just like complex subjects, adverbials in pre-subject position increase the syntactic weight of the 'object' clause in unusual, i.e. initial, position (cf. end-weight principle), making it 'nose-heavy', as it were. This, in turn, increases production effort and favours the insertion of a filler in the form of *that*.

The view of *that* functioning as a filler rather than a genuine subordinator marking the boundary between main- and subordinate clause is also supported by the following example from the corpus, where the position of *that* (if understood as a subordinator) would suggest a clause boundary after the adverbial (*in the present climate linked with her disability*). Semantically, however, the adverbial can only be understood as being part of the 'object' clause, which disqualifies *that* as a marker of subordination.

(31) *I think* in the present climate linked with her disability *that* finding a full-time tenured post will be (s1b-062-49)

The *that*-complementizer, in other words, has an important temporal function, like typical fillers, which allow the speaker to 'buy time'. This, in turn, can help alleviate production difficulties, as noted for instance by Jaeger (2005) (cf. also Clark 2004). Close analysis of the corpus data shows that there is indeed a trade-off between the use of *that* and production difficulties, with insertion of *that* correlating with fewer instances of repetition and/or restarts immediately preceding or following *I think that*. More precisely, with *that* omission we find such disfluencies in 16.7 percent (27 instances) of all cases, such as example (32). With *that* insertion, on the other hand, such disfluencies occur in only 3.4 percent (1 instance) of all cases.

(32) *I I I think* there wh some of us are in great difficulty here (s1b-028-101)

An interesting illustration of the filler function of *that* is given in (33), where *that* is abandoned in the retake, i.e. once the production difficulty has been overcome.

(33) *I think* that <,> *I think* the reality is <,> (s1a-062-162)

To sum up, the corpus results suggest that the *that*-complementizer following initial *I think* acts like a typical filler. Its function in spoken language is therefore primarily a linear one, i.e. on the temporal plane, not so much a hierarchical one, i.e. marking syntactic subordination and backgrounding. Such syntactic backgrounding, incidentally, would run counter the typical pragmatic use and information structure of these constructions, where the *that*-clause presents the main point of the message. *I think* only has secondary, qualifying function, which typically reduces the speaker's commitment to the proposition of the *that*-clause. In fact, it is precisely this hedging or distancing function of *I think* that makes the use of *that* as a marker of subordination redundant. As argued elsewhere (Kaltenböck 2006a), the hierarchical function of a *that*-subordinator (which is more prominent with CTPs of more specific semantic content) is essentially also one of distancing the speaker from the proposition it introduces. With initial *I think*, however, this distancing function is already taken care of and results in omission of *that*,[18] except where it is needed for linear purposes, i.e. as a filler.

5. Conclusion

In this paper I have tried to show that in spoken language the *that*-complementizer in object clauses no longer functions as a genuine marker of subordination (i.e. indicating syntactic hierarchy) but rather as a filler (i.e. functioning on the linear plane). This erosion of grammatical meaning of *that* can be linked to the semantic erosion of the CTP-phrase: more fully grammaticalised (pragmaticalised) CTP-phrases, involving high-frequency weak assertives, such as *I think*, *I suppose*, are no longer syntactically interpreted as main clauses, which consequently reduces the need for an overt marker of subordination. If the *that* is still used, it is usually simply a 'filling' device inserted for rhythmical purposes or to alleviate production difficulties.

Since the role of *that* is closely linked to the syntactic status of the CTP-phrase, I have first tried to show that syntactic tests intended to demonstrate a difference between CTPs followed by *that* and zero do not provide conclusive

18 A similar view has recently also been expressed by Kearns (2007: 501), who argues that "[t]he modifier sense of an epistemic verb and its subject in matrix position promotes zero in the complement clause".

evidence (Section 3.1). I have then turned to a discussion of cognitive-functional arguments, which indicate that, although epistemic CTPs are frequently downgraded, their status is largely indeterminate and depends on actual contextual realisation (Section 3.2). I have subsequently suggested prosodic realisation as a possible decisive factor for signalling foregrounding and backgrounding (i.e. main and comment clause status) of the CTP. The prosodic analysis in Section 4 has focussed on *I think* as an extreme case of grammaticalisation and shows that presence or absence of the *that*-complementizer does not correspond with different prosodic behaviour. Although *I think* + *that* reveals a slightly higher propensity to occur with a separate nuclear tone than *I think* + zero, both constructional types exhibit a similar distribution of the three prosodic patterns identified. This means that the two formal signals available for indicating relative prominence of *I think*, prosody and an explicit marker of subordination, do not match. If we take relative pitch prominence as an indication of matrix clause status, we have to conclude that both constructional variants may qualify for main clause status but at the same time very rarely do. This equivalence in actual use of the two syntactic types casts additional doubt on the subordinator function of the *that*-complementizer in spoken language, which is corroborated by prosodic evidence and co-occurrence facts (subject type, adverbials, disfluency features) suggesting that the *that*-complementizer is mainly used as a filler inserted for rhythmical reasons or to alleviate production difficulties, especially if followed by a syntactically complex ‘object’ clause.

Appendix

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	94.9% (466)	5.1% (25)	100% (491)
Public dialogue s1b (171,062)	89.3% (434)	10.7% (52)	100% (486)
Public monologue s2a (152,829)	87.9% (80)	12.1% (11)	100% (91)
Scripted speech s2b (108,164)	81.4% (57)	18.6% (13)	100% (70)
Total	91.0% (1036)	9.0% (102)	100% (1138)

Table A. Clause-initial *I think* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	96.4% (54)	3.6% (2)	100% (56)
Public dialogue s1b (171,062)	90.3% (28)	9.7% (3)	100% (31)
Public monologue s2a (152,829)	100% (5)	0% (0)	100% (5)
Scripted speech s2b (108,164)	50.0% (1)	50.0% (1)	100% (2)
Total	93.6% (88)	6.4% (6)	100% (94)

Table B. Clause-initial *I suppose* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	100% (14)	0% (0)	100% (14)
Public dialogue s1b (171,062)	75.0% (18)	25.0% (6)	100% (24)
Public monologue s2a (152,829)	83.3% (10)	16.7% (2)	100% (12)
Scripted speech s2b (108,164)	100% (2)	0% (0)	100% (2)
Total	84.6% (44)	15.4% (8)	100% (52)

Table C. Clause-initial *I hope* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	100% (1)	0% (0)	100% (1)
Public dialogue s1b (171,062)	55.6% (15)	44.4% (12)	100% (27)
Public monologue s2a (152,829)	0% (0)	100% (5)	100% (5)
Scripted speech s2b (108,164)	46.2% (6)	53.8% (7)	100% (13)
Total	47.8% (22)	52.2% (24)	100% (46)

Table D. Clause-initial *I believe* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	100% (9)	0% (0)	100% (9)
Public dialogue s1b (171,062)	85.7% (6)	14.3% (1)	100% (7)
Public monologue s2a (152,829)	100% (4)	0% (0)	100% (4)
Scripted speech s2b (108,164)	0% (0)	0% (0)	(0)
Total	95.0% (19)	5.0% (1)	100% (20)

Table E. Clause-initial *I guess* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	87.5% (7)	12.5% (1)	100% (8)
Public dialogue s1b (171,062)	83.3% (5)	16.7% (1)	100% (6)
Public monologue s2a (152,829)	100% (2)	0% (0)	100% (2)
Scripted speech s2b (108,164)	100% (1)	0% (0)	100% (1)
Total	88.2% (15)	11.8% (2)	100% (17)

Table F. Clause-initial *I'm afraid* followed by *that*- and zero in ICE-GB (raw figures in brackets)

Text type (number of words)	- *that*	+ *that*	Total
Private dialogue s1a (205,627)	100% (2)	0% (0)	100% (2)
Public dialogue s1b (171,062)	20.0% (1)	80.0% (4)	100% (5)
Public monologue s2a (152,829)	66.7% (2)	33.3% (1)	100% (3)
Scripted speech s2b (108,164)	(0)	(0)	(0)
Total	50.0% (5)	50.0% (5)	100% (10)

Table G. Clause-initial *I suspect* followed by *that*- and zero in ICE-GB (raw figures in brackets)

References

Aijmer, Karin. 1972. *Some aspects of psychological predicates in English*. Stockholm: Almqvist & Wiksell.

Aijmer, Karin. 1997. "*I think* – an English modal particle". In Swan, T.; Westvik, O. J. (eds.). *Modality in Germanic languages. Historical and comparative perspectives*. Berlin: Mouton de Gruyter, 1-47.

Asher, Nicholas. 2000. "Truth conditional discourse semantics for parentheticals". *Journal of Semantics* 17 (1), 31-50.

Biber, Douglas. 1999. "A register perspective on grammar and discourse: variability in the form and use of English complement clauses". *Discourse Studies* 1 (2), 131-150.

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan and Finegan, Edward. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Boersma, Paul & Weenink, David (2008). Praat: doing phonetics by computer (Version 4.4.33) [Computer programme]. Retrieved December 2005, from http://www.praat.org

Bolinger, Dwight. 1985. "The inherent iconism of intonation". In Haiman, John (ed.). *Iconicity in syntax*. Amsterdam: Benjamins, 97-108.

Boye, Kaspar; Harder, Peter. 2007. "Complement-taking predicates: usage and linguistic structure". *Studies in Language* 31 (3), 569-606.

Brandt, Margareta. 1984. "Subordination und Parenthese als Mittel der Informationsstrukturierung in Texten". *Sprache und Pragmatik* 32, 1-37.

Brazil, David. 1997. *The communicative value of intonation in English*. Cambridge: CUP.

Brinton, Laurel J. 1996. *Pragmatic markers in English. Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.

Brinton, Laurel J.; Traugott, Elizabeth Closs. 2005. *Lexicalization and language change*. Cambridge: CUP.

Clark, H. 2004. "Pragmatics of language performance". In Horn, L.R.; Ward, G. (eds.). *Handbook of pragmatics*. Oxford: Blackwell, 365-382.

Couper-Kuhlen, Elizabeth. 1986. *An introduction to English prosody*. London: Edward Arnold.

Cruttenden, Alan. 1997. *Intonation* (2nd ed.). Cambridge: CUP.

Davidson, D. 2001. *Inquiries into truth and interpretation* (2nd ed.). Oxford: Clarendon.

Diessel, Holger; Tomasello, Michael. 1999. "Why complement clauses do not include a that-complementizer in early child language". *Proceedings of the 25th Annual Meeting, Berkeley Linguistics Society*, 86-97.

Elsness, J. 1984. "*That* or zero? A look at the choice of object clause connective in a corpus of American English". *English Studies* 65, 519-33.

Erman, Britt; Kotsinas, Ulla-Britt. 1993. "Pragmaticalization: the case of *ba* and *you know*". *Studier I modern språkvetenskap* (New series 10, Acta Universitatis Stockholmiensis), 76-93.

Erteschik-Shir, N., Lappin, S. 1979. "Dominance and the functional explanation of island phenomena", *Theoretical Linguistics* 6 (1), 41-86.

Firbas, Jan. 1992. *Functional sentence perspective in written and spoken communication*. Cambridge: CUP.

Fischer, Olga. 2007a. *Morphosyntactic change. Functional and formal perspectives*. Oxford: OUP.

Fischer, Olga. 2007b. "The development of English parentheticals: a case of grammaticalization?". In Smit, Ute et al. (eds.) *Tracing English through time*. Wien: Braumüller, 99-114.

Givón, Talmy. 1984. *Syntax: a functional-typological introduction. Vol. 1*. Amsterdam: Benjamins.

Goodwin, Charles; Goodwin, Marjorie Harness. 1992. "Assessments and the construction of context". In Goodwin, Charles; Duranti, Alessandro (eds.). *Rethinking context*. Cambridge: CUP, 147-189.

Green, Georgia M. 1976. "Main clause phenomena in subordinate clauses", *Language* 52, 382-397.

Hand, M. 1993. "Parataxis and parentheticals". *Linguistics and Philosophy* 16, 495-507.

Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.

Hand, Michael. 1993. "Parataxis and parentheticals". *Linguistics and Philosophy* 16, 495-507.

Hooper, Joan B, Thompson, Sandra. 1973. "On the applicability of root transformations". *Linguistic Inquiry* 4 (4), 465-497.

Hooper, Joan B. 1975. "On assertive predicates" In Kimball, J.P. (ed.). *Syntax and semantics. Vol 4*. New York: Academic Press, 91-124.

Hopper, Paul J.; Traugott, Elizabeth Closs. (2003). *Grammaticalization* (2nd edition). Cambridge: CUP.

Huddleston, Rodney; Pullum, Geoffrey K. 2002. *The Cambridge grammar of the English language*. Cambridge: CUP.

Jaeger, Florian T. 2005. "Optional *that* indicates production difficulty: evidence from disfluencies". *Proceedings of DiSS'05 Disfluency in Spontaneous Speech Workshop*. 10-12 September 2005, Aix-en-Provence, France, 103-109.

Kaltenböck, Gunther. 2006a. "'...*That* is the question': complementizer omission in extraposed *that*-clauses". *English Language and Linguistics* 10 (2), 371-396.

Kaltenböck, Gunther. 2006b. "Some comments on comment clauses: a semantic classification". In Povolná, Renata; Dontcheva-Navratilova, Olga (eds.). *Discourse and interaction*. Brno: Masarykova Univserszita, 71-87.

Kaltenböck, Gunther. 2008. "Prosody and function of English comment clauses". *Folia Linguistica* 42 (1), 83-134.

Kaltenböck, Gunther. 2009a. "Initial *I think*: main or comment clause?", *Discourse and Interaction* 2 (1), 49-70.

Kaltenböck, Gunther. 2009b. "English comment clauses: position, prosody, and scope". *Arbeiten aus Anglistik und Amerikanistik* 34 (1), 49-75.

Kaltenböck, Gunther. 2009 forthc. "Pragmatic functions of parenthetical *I think*". In Kaltenböck, Gunther; Mihatsch, Wiltrud; Schneider, Stefan. (eds.). *New approaches to hedging*. Amsterdam: Elsevier.

Kärkkäinen, Elise. 2003. *Epistemic stance in English conversation*. Amsterdam: Benjamins.

Kärkkäinen, Elise. 2009 forthc. "Position and scope of epistemic phrases in planned and unplanned American English". In Kaltenböck, Gunther; Mihatsch, Wiltrud; Schneider, Stefan. (eds.). *New approaches to hedging*. Amsterdam: Elsevier.

Kearns, Kate. 2007. “Epistemic verbs and zero complementizer”. *English Language and Linguistics* 11 (3), 475-505.

Knowles, John. 1980. “The tag as a parenthetical”. *Studies in Language* 4, 370-409.

Kruisinga, E. 1932. *A handbook of present-day English*. Part II. Groningen: Noordhoff.

Langacker, Ronald W. 1991. *Foundations of cognitive grammar*. *Vol II: Descriptive applications*. Stanford: Stanford University Press.

Lepore, E. & B. Loewer (1989). “You can say *that* again”. In French, P. et al. (eds.) *Midwest studies in philosophy. Vol XIV*. Notre Dame, IN: University of Notre Dame Press, 338-56.

Linell, Per. 1998. *Approaching dialogue: talk, interaction and contexts in dialogical perspectives*. Amsterdam: Benjamins.

Mackenzie, J. Lachlan. 1984. “Communicative functions of subordination”. In Mackenzie, J.L.; Wekker, H. (eds.). *English language research: the Dutch contribution I*. Amsterdam: Free University Press, 67-84.

Mindt, Ilka. 2003. “Is *I think* a discourse marker?”. In Mengel, Ewald et al. (eds.) *Proceedings Anglistentag 2002 Bayreuth*. Trier: WVT, 473-483.

Nelson, Gerald; Wallis, Sean; Aarts, Bas. 2002. *Exploring natural language. Working with the British Component of the International Corpus of English*. Amsterdam, Philadelphia: Benjamins.

Nuyts, Jan. 2000. “Tensions between discourse structure and conceptual semantics: the syntax of epistemic modal expressions”. *Studies in Language* 23 (1), 103-135.

Peterson, Peter. 1999. “On the boundaries of syntax: non-syntagmatic relations”. In Collins, Peter; Lee, David (eds.). *The clause in English*. Amsterdam, Philadelphia: Benjamins, 229-250.

Pomerantz, Anita; Fehr, B. J. 1997. “Conversation analysis: an approach to the study of social action as sense making practices”. In Dijk, Teun A. van (ed.). *Discourse as social interaction*. London: Sage, 65-91.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; Svartvik, Jan. 1985. *A Comprehensive grammar of the English language*. Harlow: Longman.

Rohdenburg, Günther. 1998. “Clausal complementation and cognitive complexity in English”. In Neumann, Fritz-Wilhelm; Schülting, Sabine. (eds.). *Anglistentag 1998 Erfurt Proceedings*. Trier. Wissenschaftlicher Verlag, 101-111.

Ross, John R. 1973. “Slifting”. In Gross, Maurice, Halle, Morris, Schützenberger, Marcel-Paul (eds.). *The formal analysis of natural languages*. The Hague: Mouton, 133-169.

Sadock, Jerrold M. 1984. “The pragmatics of subordination”. In Geest, Wim de; Putseys, Y. (eds.). *Sentential complementation*. Dordrecht: Foris, 205-213.

Schegloff, Emanuel. 1990. “On the organization of sequences as a source of ‘coherence’ in talk-in-interaction”. In Dorval, B. (ed.). Conversational organization and its development. Norwood, NJ: Ablex, 51-77.

Simon-Vandenbergen, Anne-Marie. 2000. “The functions of *I think* in political discourse”. *International Journal of Applied Linguistics* 10 (1), 41-63.

Stenström, Anna-Brita. 1994. *An introduction to spoken interaction*. London: Longman.

Stenström, Anna-Brita. 1995. “Some remarks on comment clauses”. In Aarts, Bas; Meyer, Charles F. (eds.). *The verb in contemporary English*. Cambridge: CUP, 290-299.

Svensson, Jan. 1976. “Reportindicators and other parentheticals”. In Karlsson, F. (ed.). *Papers from the Third Scandinavian Conference of Linguistics*. Turku: Textlinguistics Research Group, Academy of Finland, 369-380.

Tagliamonte, Sali; Smith, Jennifer. 2005. “No momentary fancy! The zero ‘complementizer’ in English dialects”. *English Language and Linguistics* 9 (2), 289-309.

Thompson, Sandra A. 2002. “’Object complements’ and conversation. Towards a realistic account”. *Studies in Language* 26 (1), 125-164.

Thompson, Sandra A.; Mulac, Anthony. 1991. “The discourse conditions for the use of the complementizer *that* in conversational English”. *Journal of Pragmatics* 15, 237-251.

Tomlin, Russell. 1985. “Foreground-background information and the syntax of subordination”. *Text* 5, 85-122.

Traugott, Elizabeth Closs. 1995. “Subjectification in grammaticalisation”. In Stein, Dieter; Wright, Susan (eds.). *Subjectivity and subjectivisation*. Cambridge. CUP, 31-54.

Urmson, J. O. 1952. “Parenthetical verbs”. *Mind* 61, 480-496.

Van Bogaert, Julie. 2006. “*I guess, I suppose* and *I believe* as pragmatic markers: garammaticalization and functions”. *Belgian Journal of English Language and Literatures* 4, 129-149.

Wells. John C. 2006. *English intonation: an introduction*. Cambridge: CUP.

Wichmann, Anne. 2000. *Intonation in text and discourse. Beginnings, middles and ends*. Harlow: Longman.

Wichmann, Anne. 2001. “Spoken parentheticals”. In Aijmer, K. (ed.). *A wealth of English. Studies in honour of Göran Kjellmer*. Göteborg: Acta Universitatis Gothoburgiensis, 177-193.

Decoding sounds: an experimental approach to intelligibility in ELF

*Ruth Osimk, Vienna**

1. Introduction

The notion of intelligibility is a highly complex matter which is thought to consist of a great number of factors. This particularly holds true for situations in which the language of communication is a non-native language for all participants, in other words a *lingua franca* (Seidlhofer 2001: 146). English has come to be the world's global language, with an estimated number of 300-400 million second and approximately 500-700 foreign language users (Crystal 2000: 10). Therefore, the amount of speakers who use English as a second or foreign language clearly exceeds the estimated number of 350-450 million first language users (Crystal 2000: 9). Moreover, it is assumed that approximately 80% of all communication occurs in the absence of native speakers (cf. Carter 1998). For this reason, the question of what hinders or promotes intelligibility in such communicative situations is a crucial one.

One ground-breaking approach to this issue was Jenkins' (2000) empirical study of communication breakdown in naturally occurring conversations between non-native speakers (NNS) of English. Furthermore, a considerable number of studies exploring factors involved in intelligibility have been carried out in psycholinguistics and also acoustic-phonetics (henceforth referred to as *intelligibility studies*). Naturally, these approaches have differed considerably from Jenkins' approach, not only with regard to methodology but also to their various underlying assumptions (cf. 3.3.). Moreover, intelligibility studies have rarely concerned themselves with the intelligibility of NNS to non-native listeners (NNL)[1] and, to the author's knowledge, no extensive psycholinguistic studies on the role of segmentals for intelligibility

* The author's email for correspondence: ruth.osimk@univie.ac.at.

1 Studies which have investigated intelligibility between NNS and NNL include Bent & Bradlow (2003), Florentine (1985), Fayer & Krasinski (1987: 313), Field (2005), Smith & Bisazza (1982), Munro, Derwing & Morton (2006).

in an English as a *lingua franca* (ELF) context have been carried out. Therefore, this paper sets out to explore this topic on a segmental level, using methods commonly applied in psycholinguistics (dictation method) and acoustic-phonetic analysis[2] (measurement of length of Voice Onset Time). This method analyses segmentals in their phrasal and syntactic co-textual but not contextual environment (cf. Widdowson 2004: 59-73 for a detailed discussion of the terms *co-text* and *context*). One should, therefore, be wary of direct comparisons to empirical approaches, such as Jenkins' observations, which are compiled from naturally occurring conversations.

In section 2 of this paper, I clarify the working definition of the term intelligibility and illustrate the dictation method used in the pilot study. Section 3 provides a brief overview of various factors which influence intelligibility. Moreover, it explains which deductions can be drawn from the degree of influence on intelligibility of the speaker, listener and item-related factors. I also address the role of phonetics and phonology for intelligibility. Finally in section 3, I also summarise the findings of Jenkins' Lingua Franca Core (LFC) and touch on underlying assumptions in the intelligibility studies that can be problematic when applied to an ELF context. In section 4, I introduce the pilot study. This was conducted to explore the importance of aspiration, different realisations of the interdental fricative and /r/ for intelligibility in ELF.

2. Defining intelligibility

2.1. What is intelligibility?

There have been various approaches to defining the term *intelligibility*, not all of which can be discussed in this paper. As a working definition for present purposes, Derwing & Munro's (1997) sub-categorisation into subjective and objective intelligibility was decided to be most suitable. 'Objective' intelligibility is defined as "the extent to which a speaker's utterance is actually understood" (Munro, Derwing & Morton 2006: 112), whereas subjective intelligibility (also referred to as *comprehensibility*) is seen as the "listeners' estimation of difficulty in understanding the message" (Munro, Derwing & Morton 2006:112).

---

[2] This paper is based on the author's MA thesis *Aspiration, [θ]/[ð] und /r/ in Englisch als Lingua Franca – eine psycholinguistische Studie zu drei Vorschlägen des Lingua Franca Core*, written at the Department of General and Applied Linguistics under the supervision of Prof. Dr. Wolfgang U. Dressler.

Derwing & Munro (1997), Munro & Derwing (1995a), Munro, Derwing & Morton (2006) also clearly differentiate intelligibility and comprehensibility from *accentedness*, which is defined as "the degree to which the pronunciation of an utterance sounds different from an expected production pattern" (Munro, Derwing & Morton 2006: 112). These three terms are seen as "related but partially independent dimensions" (Derwing & Munro 1997: 2). It was found that "although some features of accent may be highly salient, they do not necessarily interfere with intelligibility" (Derwing & Munro 1997: 11). Their results showed that accent sometimes had a negative effect on intelligibility but that this effect did not correlate with the degree of accent and that even strong accents did not necessarily result in poor intelligibility (cf. Munro & Derwing 1995a: 301, 1995b: 74; Munro 1998: 139ff). It therefore seems necessary to differentiate between these three terms, especially when applied to an ELF context, as in the pilot study in this paper.

2.2. Dictation method

In the pilot study described in this paper, the dictation method was used to assess the intelligibility of different realisations of the tested segmentals. This is a common method to assess objective intelligibility (used by e.g. Brodkey 1972, Derwing & Munro 1997, Bent & Bradlow 2003, Burda et al. 2003), in which the listeners hear spoken or read utterances and are asked to transliterate them. Moreover, Brodkey (1972: 205) found that the dictation method was useful for testing "mutual intelligibility of neighbouring speech communities or dialectal groups" and could identify loss of information between persons who "ostensibly speak the same language" (1972: 216). This suggests that the dictation method is also likely to be useful in testing intelligibility in NNS-NNL conversations.

An obvious disadvantage of the dictation method is that the context of situation is not taken into account and that understanding individual words does not imply a general understanding (cf. Zielinski 2004). Nevertheless, it has found to be "a useful window on the listeners' comprehension" (Munro, Derwing & Morton 2006: 113). Furthermore, Munro, Derwing & Morton (2006: 113) observe a correlation between listener intelligibility judgements via the dictation method and the "phonological properties of the speaker's output". As it can therefore be expected that the results reflect the degree of influence of segmentals on intelligibility, the dictation method was considered useful for the tests in this study.

3. Factors affecting intelligibility

A wide range of research has been carried out on the topic of intelligibility and second language acquisition. The topic has been approached from socio- and psycholinguistic and also acoustic-phonetic perspectives. However, despite the extensive body of research conducted on the issue, there is little common ground on which to compare the results. The studies vary greatly with regard to combination of native and non-native speakers and listeners, first languages of the test persons, methods and even their definition of the term *intelligibility*.

3.1. Talker, listener and item-related factors

Intelligibility studies generally investigate which factors influence intelligibility and to what extent. Bradlow & Pisoni (1999: 2074) divide these factors into three categories (a similar sub-categorisation can be found in Munro, Derwing & Morton 2006: 114 and in Bent & Bradlow 2003: 1600), namely talker-, listener and item-related factors. Talker-related factors are ways in which the speaker chooses to adapt to the challenges of a situation, e.g. through clear speech, and adaptation of volume and speed (Bradlow & Pisoni 1999: 2074). Listener-related factors refer to the influence of familiarity with individual speakers, NNS in general, different accents and the effect of the topic of conversation on the comprehension-process. There is also a broad empirical basis attesting that the listeners' attitudes (e.g. annoyance) sometimes influence intelligibility judgements (e.g. Einstein & Verdi 1985).[3] Item-related factors, finally, refer to all aspects concerning the properties of the language-input of the conversation partner itself and the influence of linguistic subsystems, i.e. syntax, lexicon, phonetics and phonology (Magen 1998: 382; Munro 1998: 151), and noise.[4]

The extent to which talker, listener and item-related factors contribute to intelligibility is crucial with regard to the conclusions that can be drawn for a given practical application, e.g. language teaching. If listener-related factors are of importance, teachers' judgements of what is intelligible will be less reliable because they might be influenced by their familiarity with the students' accents (Brodkey 1972: 216). A higher significance of item-related factors would indicate a certain universality of linguistic properties which facilitate intelligibility. In this case, important conclusions could be drawn as

[3] Cf. also Niedzielski & Preston (2000) for the influence of language attitudes on perception.

[4] For an extensive literature review cf. Osimk (2007: 33-44).

to which aspects to prioritise in pronunciation teaching. Regarding the pilot study, a high consensus for listeners with similar language backgrounds and experiences would indicate a stronger influence of listener-related factors. However, if listener groups – regardless of their language experience – concurred as to which language properties are easy to understand, this would point to a greater importance of item-related factors (Hazan & Markham 2004: 3109; Munro, Derwing & Morton 2006: 114).

The importance of item-related factors over the listener-related factors was suggested by a number of studies. In an early study, Flege (1988) noticed that the listeners made similar intelligibility judgements concerning accentedness. Similarly, Smith & Bisazza (1982) observed that the listeners agreed as to which speakers were most difficult to understand and reported that "89% of the subjects responded that the Indian speaker was the most difficult to comprehend" (1982: 267). Later studies gave similar findings: Munro, Derwing & Munro, for example, showed that two listener groups, English native listeners (ENL) and NNL with different first languages, largely agreed on "which of the 48 speakers were the easiest and most difficult to understand; between-group effect sizes were generally small" (2006: 111). They conclude that the listeners' listening experience contributed less to understanding than item-related factors (2006: 125). Hazan and Markham (2004) describe the correlation between NL children and adults as to which of the NS were the easiest or most difficult to understand as a "striking fact" (2004: 3112). Finally, Major et al.'s (2002) findings show a strong tendency that the listeners agree on which accents were rather intelligible or unintelligible.

3.2. Role of phonetics and phonology for intelligibility

In various studies, phonetics and phonology have been shown to play an important role for intelligibility (e.g. Fayer & Krasinski 1987; Anderson-Hsieh, Johnson & Koehler 1992; Magen 1998). There is evidence that suprasegmentals contribute considerably to intelligibility. However, to the author's knowledge, only a small number of studies have been carried out which have investigated this for NNL. Two examples include Jenkins' (2000) and Field's (2005) observations, which indicate the importance of nuclear stress for NNL understanding. Jenkins (2002: 89) observed in her corpus that "misplaced tonic (nuclear) stress along with a consonant substitution within the wrongly stressed word" sometimes led to miscommunication and that this was also the case in situations where "tonic stress [was] misplaced on words that are both familiar to listeners and contain no segmental errors" (ibid.).

However, other studies have also stressed the importance of segmentals for intelligibility. Fayer & Krasinksy (1987), for example, tested NN-speech for NL and NNL and found not only that pronunciation seemed to contribute more to intelligibility than all other language sub-systems, but also that pronunciation appeared to be of particular importance and that "[i]ntonation, word choice and voice trail behind" (1987: 322). Only very few studies investigate what helps or hinders intelligibility on a segmental level. Some studies highlight that the change of a segment, as well as elision or the addition of segments, can lead to perception errors (e.g. Anderson-Hsieh, Johnson & Koehler 1992: 544, Bradlow & Pisoni 1999: 2084, Munro 1998: 150). However, these observations regarding segmentals are usually not discussed in any detail.

3.3. Jenkins' (2000) Lingua Franca Core

Two extensive studies that have been carried out on the role of segmentals in NNS-intelligibility are Jenkins' (2000) corpus and Hirschfeld's (1994) study for L1 (first language) German listeners.[5] Jenkins' aim was to investigate which phonological features or L1-based phonetic variations (cf. Seidlhofer 2004: 216) most frequently caused communication breakdown or intelligibility problems in ELF communication. In one part of the investigation, Jenkins observed forty instances of communication breakdown, whereby the majority (27 out of 40) were caused by 'errors' on the phonetic and phonological level (2000: 85). The remainder were due to lexical and grammatical deviations, world knowledge and ambiguities. The phonetic and phonological features which were found crucial for successful communication in ELF were summarised in the LFC (Jenkins 2000). This empirical corpus differs from most of the intelligibility studies in various aspects, especially with regard to method and extent. The data was collected over the duration of three years and through observation and documentation of "genuine interactional speech data" (Jenkins 2000: 131). In contrast, in most experimental intelligibility studies, data was secured in a short period of time and in controlled settings.

Moreover, Jenkins' observations are based on certain underlying assumptions, some of which differ to a large extent to those in experimentally

[5] Hirschfeld's and Jenkins' findings were similar with regard to the importance of aspiration, vowel quantity over quality and the problematic nature of alternating pronunciation of the central vowel (cf. Osimk 2007 for a more detailed discussion of the two investigations).

conducted intelligibility studies.[6] Most importantly, Jenkins considers ELF as a phenomenon in its own right, and not merely as an inferior variant of ENL. Even though the LFC is based on L1-varieties, Jenkins does not contrast ENL and ELF in a judgemental way and views non-native utterances, if intelligible, as "perfectly acceptable instances of L2 sociolinguistic variation" (Seidlhofer 2004: 217). Contrary to a large number other intelligibility studies, "a genuine difference perspective" as opposed to "a deficit [...] perspective" (Seidlhofer 2004: 217) is assumed, on the basis of demonstrable irrelevance of certain features.

The deficit perspective addressed by Seidlhofer (2004) is apparent in the majority of experimental intelligibility studies with NNL. It manifests itself in four underlying assumptions, which are problematic regarding intelligibility from the position of the NNL and especially for ELF communication. These four assumptions are that NS are more suitable than NNS to judge which factors influence intelligibility; the use of NS approximation as primary goal in pronunciation teaching; the supposition that accentedness equals poor intelligibility; and that NNS-communication is, per se, less successful than "purely 'native' speech communication" (van Wijngaarden, Steeneken & Hourgast 2002: 1906).[7] However, in the light of the expansion of English as a global language these assumptions are in urgent need of reconsideration and re-evaluation, especially as the growing body of ELF-research shows that NNS do not, by any means, communicate unsuccessfully but rather highly effectively in ELF situations (e.g. Firth 1996, Meierkord 1996). Unfortunately, a detailed, critical discussion of the application of these four underlying assumptions is beyond the scope of this paper, but can be found in Osimk (2007) or Rajadurai (2007).

4. Pilot study

This section introduces the pilot study that was conducted to assess intelligibility from a segmental point of view, assessing the importance of aspiration and of the different realisations of the interdental fricative and /r/ for intelligibility.

---

6 Cf. Osimk (2007: 53ff) for a more detailed discussion.

7 For an extensive discussion of these assumptions and their implications cf. Osimk (2007) or Rajadurai (2007).

4.1. Aims

The aim of this pilot study was to test three phonological features with psycholinguistic methods and to investigate how these relate to Jenkins' results. The three aspects (aspiration, realisations of [θ]/[ð] and /r/) were chosen for a number of reasons. Firstly, aspiration is regarded as playing a crucial role for intelligibility by two extensive studies, Jenkins (2000) and Hirschfeld (1994). The interdental fricative as in *thing* [θ] und lenis in *that* [ð] is a 'typical' English sound that receives much attention in English language teaching (ELT). However, according to Jenkins (2000), it does not cause problems for intelligibility, apart from when pronounced [s]/[z]. The third phonological feature, /r/, was chosen as its rhotic variant was included in the LFC, not because it necessarily eased intelligibility but because greater teachability and reduced redundancy was assumed (Jenkins 2000). The realisations were also compared to the standard pronunciation[8] of the phonemes.

Moreover, the aim of the pilot study was to explore the importance of syntactic and phrasal co-text, as well as the severity of the listener versus item-related factors, for intelligibility. For listener-related aspects, the influence of different types of familiarity on the intelligibility scores was investigated. These were familiarity with the utterances produced by the listeners' own accents (i.e. substitutions which are commonly made by EFL-learners of the listeners' L1 as well as sounds from their own L1) and familiarity with other non-native accents through previous experience. The latter was determined by means of a questionnaire completed by all listeners.

4.2. Preparation of stimuli

The stimuli were taken from the online resource *Speech Accent Archive* (SAA),[9] a collection of recordings of the same English text, read by a large number of ENS and NNS. The text has a length of 69 words, uses frequent vocabulary and contains a large number of examples of the English phonemic

8 The author is aware that standard language ideology is a highly complex and controversial issue. However, for the purpose of this paper, the term *standard* was chosen to refer to Received Pronunciation (RP) and General American (GA). This paper is directed, to a large degree, towards pedagogical applications and RP and GA are the varieties most commonly taught in ESL/ESP (Jenkins 2003: 31). One of the aims of the pilot study was to compare GA/RP standard pronunciations to alternating realisations to determine which realisations were most intelligible. For this reason, the issues of standardisation and ownership of English are not discussed in detail.

9 The SAA was used for this study with the kind permission of Steven H. Weinberger.

inventory. The large number of speakers and a search function for specific realisations, (e.g. *non-aspiration*) made the SAA a suitable tool for stimuli selection for this pilot study. However, the shortness of the text, and thus the limited number of test words available, are an unavoidable disadvantage for the reliability of the study. The chosen realisations were segmented with the programme STx[10] and saved as mono files. In order to gain a larger number of stimulus words, two sets of data were created, whereby the same realisations occurred in both sets of data and only the speakers and words varied.

4.3. Methods

The study was conducted in two parts using the dictation method. In test 1, the stimulus words were tested in isolation, while test 2 assessed the same stimuli in their phrasal co-text or with their syntactic constituents. To prevent effects of familiarity, the listeners only participated either in test 1 or in test 2 of the study.

4.4. Participants

4.4.1. Speakers

The readings of the text for 13 speakers, 8 male and 5 female, aged 18 to 66 were chosen from the SAA. The distribution of first languages was 3 Spanish, 2 Italian, 4 French and 4 German (3 German, 1 Swiss German). For the pilot study, the speakers were chosen according to their first languages and according to how many of the tested features (aspiration and different realisations of the interdental fricative and /r/) were produced while reading the text. The distribution of countries of origin was highly diverse; for example, the three Spanish speakers came from Venezuela, Nicaragua and Spain. It can not be excluded that these accent variations influenced the intelligibility scores, e.g. regarding the experience of the listeners with these accents.

10 The programme STx (Version 3.7.5) was provided by the Vienna Acoustics Research Institute and used with their permission.

4.4.2. Listeners

In total, 64 listeners aged 19-31, 23 male and 33 female, with the L1s French, German, Italian and Spanish participated in the study. Most of the participants were exchange students at the University of Vienna at the time the study was conducted, and therefore, their level of education can be assumed to be fairly similar. All but one participant (sp13) came from European countries, i.e. depending on L1, from Spain, Italy, France or Austria. The participants indicated that they had grown up monolingually, with the exception of the 5 Spanish speakers (4 were bilingual in Spanish/Catalan and 1 in Spanish/Galician), and one participant who was bilingual in French/Hungarian.

From the years of language instruction and experience which the listeners had indicated in the questionnaire, all listeners can be regarded as advanced learners of English. The exchange students were considered suitable for this study, as they were assumed to be regular ELF users in their daily lives (and this was also confirmed by the questionnaire answers). At the same time, however, the participants represent a specific target group with similar ages and educational backgrounds which may or may not be representative of other segments of the population (e.g. who differ in their educational background).

4.5. Procedure

The participants listened to the words in isolation (test 1) or in their syntactic and phrasal co-text (test 2) and were asked to transliterate what they understood. If needed, it was possible for the participants to listen to the stimuli a second time. Furthermore, the listeners were requested to note down any comment they wished to add. For every L1, 8 listeners participated in test 1 (total number =32). The order in which the stimuli were played varied. For test 1, the listeners heard 38 words distributed as follows: 7 words for aspiration, 9 words with variations of the interdental fricative and 11 variations of /r/. Additionally, 10 ‘dummy’ words, which contained none of the tested features, were played in between to minimise an effect of familiarity with regard to the tested features.

In test 2, the words from test 1 were examined in context. As the main focus was on investigating the intelligibility of the words in isolation, only 24 listeners were tested, 6 for each of the 4 first languages. For each feature, the text was divided into 12 parts (cf. 8.2.), leaving gaps for the words containing this feature. Each feature was tested separately and the listeners were only tested on one of the features to avoid familiarity effects. In testing the interdental fricative, for example, the listeners heard phrase (1) and were

asked to fill in the missing words (*snack* – dummy word and *brother* – interdental fricative).

(1) and maybe a _____ for her _____ Bob.

For both tests 1 and 2, the participants were presented with the stimuli from only one of the data sets. Additionally, the listeners were asked to fill in a questionnaire containing questions on their language experience, familiarity with other accents, age, learning method etc.

In sections 4.6 to 4.8, the background and method of each of the tested features (aspiration, different realisations of the interdental fricative and /r/) is described, followed by an analysis of the results of the pilot study.

4.6. Aspiration

4.6.1. Background and Method

Aspiration was tested by measuring the length of Voice Onset Time (VOT) with an acoustic-phonetic analysis. VOT has been defined as the "interval between the release of an articulatory gesture, usually […] a stop and the beginning of vocal fold vibration" (Cho & Ladefoged 1999: 225) and can been seen as an indicator for aspiration, whereby aspirated stops have a longer VOT than unaspirated stops.[11] VOT is influenced by various factors. The most important observation is that the VOT is longer the further back it is articulated, e.g VOT for velar stops is longer than for alveolar or labial plosives (Docherty 1992: 25; Cho & Ladefoged 1999: 208, Yao 2007: 185).

Languages differ in their categorisation of plosives, as well as in which VOT-length they attribute to aspirated and unaspirated plosives (Cho & Ladefoged 1999: 223). English distinguishes two categories of plosives, namely unaspirated and usually unvoiced (Ladefoged 2005: 137) /b d g/ and /p t k/ which are aspirated in word- and syllable-initial position (Khattab 2000: 95) and fortis unaspirated in consonant clusters /sp/, /st/, /sk/, as well as when followed by /s/. A number of studies have measured VOT for English and the results vary to some degree. This variability is likely to be due to

11 VOT can be seen as indicator for aspiration rather than voicing, as there are languages with voiceless, unaspirated plosives, e.g. Austrian German, where lenis plosives are not voiced but have a shorter VOT than fortis plosives. Moreover, VOT is only an indicator for languages with aspirated, voiced plosives (e.g. Hindi) which have a long VOT but also vocal fold vibration (comment by Sylvia Moosmüller, personal communication, 2007)

differences in the context in which the words were tested and the tested variety of English, as well as the fact that the border between voiced and voiceless plosives is assumed to be a continuum. For the purposes of this study, it was useful to consider the variability of the VOT-measures and to divide the plosives into 0-20ms (unaspirated) and 40-60ms (aspirated) and to assume a transition area between 20-40ms.

The listeners' first languages differ in the way they subcategorise their plosives. Germanic languages, such as German, contrast between aspirated and unaspirated voiceless plosives. Therefore, there is no vocal fold vibration in these languages. Romance languages (French, Italian and Spanish in this study) however, differentiate between voiceless-unaspirated and voiced plosives (Ladefoged 2005: 137). While the English velar plosive /k/ has a VOT of about 50-60ms, the VOT for the same sound in Spanish is only about 20ms. It could, therefore, be expected that listeners with a Romance L1 will recognise fortis plosives with a shorter VOT more easily than listeners with the L1 German.

For the study, VOT was measured with spectrograms and waveforms, using the programme STx. The measurement was taken from the plosive release until the start of the vocal fold vibration of the following vowel, the first positive zero-crossing. The eleven plosives were first analysed according to recognised words and then according to recognised feature. The feature /k/, for example, was counted as recognised when the plosive was identified by the listener, e.g. when *call* instead of the target word *car* was transliterated.

4.6.2. Results

The analysis shows the clear tendency that words containing fortis plosives with a longer VOT were recognised more often than those with a shorter VOT. Words with a VOT of 0-20ms were recognised 5.7 times on average (n=16), words with a VOT of 40-60ms 12.7 times (p=0.05) (cf. Figure 1). Despite the margin of error involved, a clear enhancement in intelligibility for plosives with a longer VOT could be shown.

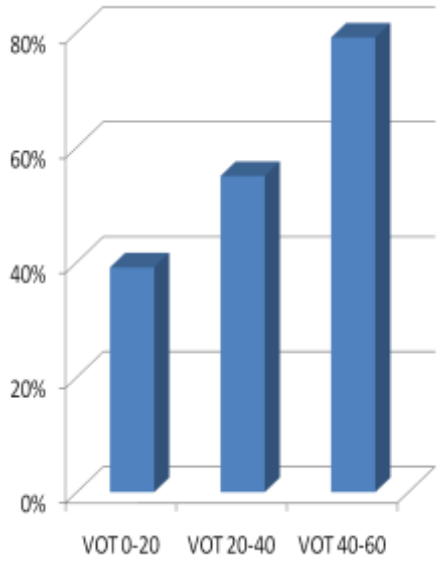


Figure 1: Correctly recognised features according to length of VOT in percent: VOT 0-20ms=39.1, 20-40ms =55, 40-60ms= 79.1

No advantage was observed when the fortis plosives were realised with the VOT which is standard for the listeners' L1. The plosives with a shorter VOT were not recognised any more often by the Spanish, Italian and French listener groups than by the listeners with L1 German, where the boundary between voiceless and voices is similar to that of English (cf. Ladefoged 2005: 137).

4.7. Interdental fricative

4.7.1. Background and Method

The English interdental fricative is often a challenge for foreign language learners as it is not part of the phonemic inventory of many languages. English differentiates between the lenis [ð], as in the tested words *the*, *with*, *brother* and *these*, and the fortis [θ], as it occurs in the tested words *things*, *thick* und *three*. In the listeners' first languages, dental fricatives only occur in Spanish, for example in the words *dedo* [ˈdeðo] or ciudad [θjuˈðað]. Therefore, the interdental fricative is often substituted by NNS with /s z/, /t d/ or /f v/. If the listeners' L1 had a considerable influence on intelligibility, the listeners would, therefore, be expected to better understand substitutions which are commonly made by L2-speakers of their own L1. Moreover, it could be expected that the Spanish listeners would reach higher intelligibility

rates for 'unsubstituted' [ð] [θ], as this is also part of their L1 inventory. Also of interest is whether one particular substitution caused greater problems than others and whether this was the same for all listener groups.

For the purpose of this investigation, the words were divided into those containing the voiced variant and those containing the voiceless variant. What followed from this division was that there were only a small number of words per variant that could be tested. Moreover, [θ] always occurred in word-initial position in the tested words whereas [ð] also occurred in medial (*brother)* and word-final (*with*) position. Because of the already small number of stimulus words, no division according to syllable position was made. As this can affect aspects such as voicing, however (as English word-final lenis consonants are partly devoiced), it cannot be ruled out that this might have influenced the results.

4.7.2. Results

Both for the lenis and the fortis interdental fricative it was found that the lowest intelligibility scores were attained if substituted by /s/ and /z/. This is observed for both the percentage of recognised words, as well as the recognised feature (cf. Table 1).

	Words	% Correctly recognised Words	% Correctly recognised F	Total number of words
[ð] as /s z/	these	0.0%	18.8%	16
[ð] as /t d/	the, brother, these (2)	34.8%	60.7%	112
[ð] as /f v/	with (2)	43.8%	43.8%	32
[θ] as /s z/	things, three (2)	41.7%	47.9%	48
[θ] as /t d/	things	68.8%	100.0%	16
[θ] as /f v/	things (2)	65.6%	96.9%	32
[θ] norm	thick, things	56.3%	71.9%	64

Table 1: Percentages of correctly recognised words and features (F) of [ð]and [θ] .(2) = this word was produced by two speakers, spoken with the same variant. Total number of words = the total number of times words containing this variant were listened to.

For all listener groups, [θ] was understood well in all realisations except as an alveolar fricative (cf. Table 2). [ð] realised as a labiodental fricative reached slightly lower intelligibility scores with the Spanish listeners than with the other listener groups. Apart from the French listeners, 50% of whom

recognised /z/ as [ð] in *these*, all listener groups reached low intelligibility scores for this realisation. [θ] realised as /s z/ in the words *things* and *three* was better understood by the French and German-speaking listener groups than it was by the Spanish and Italian listener groups. In addition, the intelligibility scores for *these* (with [ð] realised as /t/ were low for all listener groups. The reason for this may lie in the fact that the dental fricative was pronounced fortis and not lenis, as commonly pronounced in standard English.

Feature	Number of words containing this realisation	Word(s)	Spanish	Italian	French	German
[θ] as /s z/	3	things, three	25,0%	41,7%	50,0%	75,0%
[ð] as z/	1	these	25,0%	0,0%	50,0%	0,0%
[θ] as /t d/	1	things	100,0%	100,0%	100,0%	100,0%
[ð] as /t d/	4	the, brother, these	57,1%	64,3%	60,7%	57,1%
[θ] as /f v/	2	things	87,5%	87,5%	100,0%	100,0%
[ð] as /f v/	2	with	25,0%	50,0%	50,0%	50,0%
[θ] as [θ]	4	thick, things	75,0%	75,0%	50,0%	75,0%

Table 2. Percentage of recognised features [θ] and [ð] of the test words according to listener groups.

Observations concerning the substitutions common in the listeners' own first language were not consistent across all listener groups. Only for the feature [θ], realised as /s z/ in *things* and *three*, can it be said that the French and German listener groups, for whom this substitution is a common one, understood these better than the Spanish and Italians. However, [ð] pronounced /s z/ (*these*) was advantageous for the French but not for the German listener group. It could also be shown that the Spanish did not profit from the norm-pronunciation any more than the other listener groups. The norm was generally well understood. Also, it was observed that the substitutions were either rather well or not well understood, regardless of listeners' L1s.

4.8. Variations of /r/

4.8.1. Background and method

The realisation of /r/ in different languages is highly diverse. It can be produced as trill, flap, tap, fricative or approximant and can be alveolar, coronal, dorsal or uvular (Ladefoged & Maddieson 1996: 214). Ladefoged & Maddieson propose that "the overall unity of the group seems to rest mostly on historical connections between these subgroups, and on the choice of the letter /r/ to represent them all" (1996: 245). In English, /r/ is usually realised either as an alveolar (BrE) or a retroflex (GA) approximant. Other realisations also occur, e.g. [ɾ], [ʀ] und [r] (Foulkes & Docherty 2001: 27). Moreover, a distinction between rhotic and non-rhotic varieties is made.

One aim of this study was to test how rhotic and non-rhotic realisations influence the intelligibility of English in a lingua franca context and which of the variants of /r/ was the most intelligible for the four listener groups. The three tested variants were alveolar, uvular and retroflex (norm) /r/. However, given the limited number of available stimuli, only place and not manner of articulation (trill, fricative, flat/tap) could be considered and syllable position was not taken into account. It cannot be excluded that these aspects equally influenced the intelligibility scores.

4.8.2. Results

The total number of recognised words showed a marked difference between rhotic and non-rhotic realisation, whereby non-rhotic /r/ produced higher intelligibility scores (rhotic 11.5%, n=5 vs. non-rhotic 58.3%, n=2). This advantage held true for all listener groups (see Table 3). Compared to the total score of recognised feature /r/, however, this was not the case (rhotic 64.6% vs. non-rhotic 70.8%). A difference was only observed for the L1 German listener group. The recognition of the feature /r/ for this group was 53.1% for rhotic realisation and 83.3% for non-rhotic realisation. For all other listener groups, the difference of recognition for the two realisations was less than 10%. (cf. Table 4).

Words	Spanish	Italian	French	German
rhotic	9.4%	21.9%	15.6%	12.5%
non-rhotic	41.7%	58.3%	58.3%	75.0%

Table 3: Comparison of recognised words with rhotic and non-rhotic variants, according to listeners' L1.

Feature	Spanish	Italian	French	German
rhotic	71.9%	59.4%	65.6%	53.1%
non-rhotic	66.7%	66.7%	66.7%	83.3%

Table 4: Comparison of recognised feature with rhotic and non-rhotic variants, according to listeners' L1.

For all listeners taken together, no significant difference was observed for the realisation of /r/ as an alveolar, uvular or standard variant. This held true both for the number of correctly recognised words, as well as the recognition of the feature /r/. There was a visible tendency however, which showed that the uvular realisation received slightly lower intelligibility scores than the alveolar and standard pronunciation. For recognition of the feature /r/, there was only an insignificant difference between alveolar (70.8%) and standard pronunciation (76.3%). For the uvular realisation, feature /r/ was recognised only 51.1% of the time.

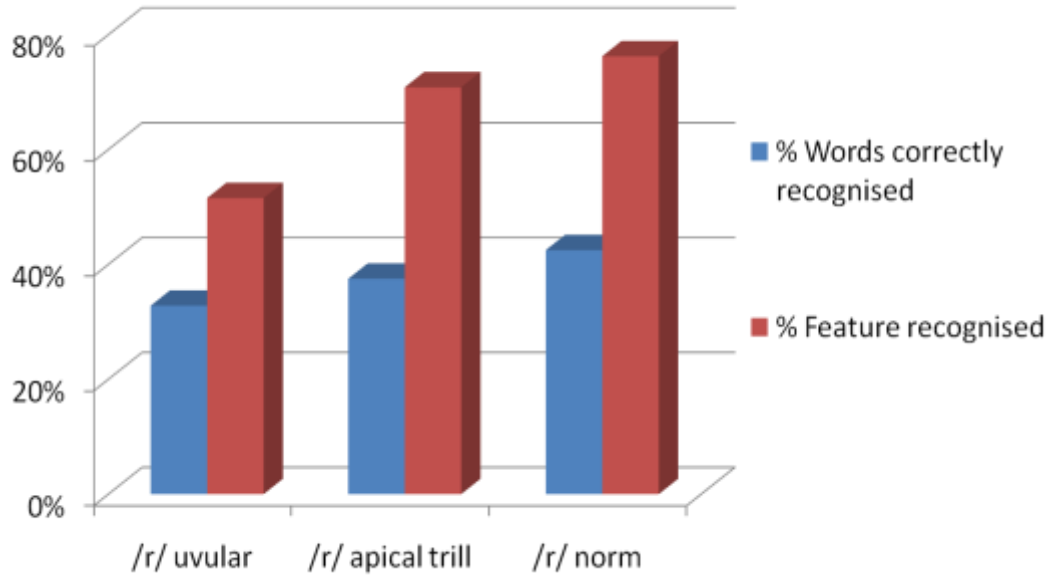


Figure 2: Comparison of correctly recognised words and feature /r/ for all listener groups, according to place of articulation. /r/ uvular: words 32.8%, feature 51.6%; /r/ apical trill: words 37.5%, feature 70.8%, /r/ norm words 42.5%, feature 76.3%.

An advantage for all groups except the German-speaking group was observed for the realisations which occur commonly in the listeners' L1s. While for Spanish and Italian, alveolar /r/ is common, the common realisation in French and German is uvular /r/. The Spanish speakers attained higher intelligibility scores for an alveolar realised /r/ produced by Spanish speakers than by Italians. The same held true for Italian listener groups. Equally, the French

benefited slightly when uvular /r/ was produced by French speakers compared to when produced by German speakers. This was non-beneficial only to the German-speaking group when hearing uvular /r/ produced by other German speakers. They did attain higher scores for uvular /r/ produced by the French speakers (cf. Table 5). This advantage was, however, merely a tendency and would need to be investigated further in order to be able to draw more meaningful conclusions.

Feature recognised according to L1	Spanish listeners	Italian listeners	French listeners	German listeners
alv. /r/ Spanish speakers	75.0%	65.0%		
alv. /r/ Italian speakers	66.7%	75.0%		
uvl /r/ French speakers			75.0%	45.0%
uvl /r/ German speakers			68.8%	37.5%

Table 5: Recognised feature /r/ according to common variant in listeners' L1.

4.10. Test 2 – phrasal and syntactic co-text

Comparing the intelligibility scores for the isolated words in test 1 and the same words in phrasal and syntactic co-text in test 2, a clear advantage for the words heard in co-text was evident (cf. Figure 3). This held true for all listener groups and manifested itself in two ways: firstly, the words were recognised more often in co-text and secondly, the words which were not recognised were interpreted differently to those heard in isolation. The word *into*, for example, heard in isolation was interpreted as *indoor*, *industry*, *injured* or *indo.* In context _____ *three red bags,* however, (gap as in test2) *into* was interpreted as *on*, *in the* or *twenty*. The listeners, therefore, adapted the stimuli to the phrasal and syntactic co-text and also to a greater degree (*on*, *twenty*), in order to place the words in a meaningful context.

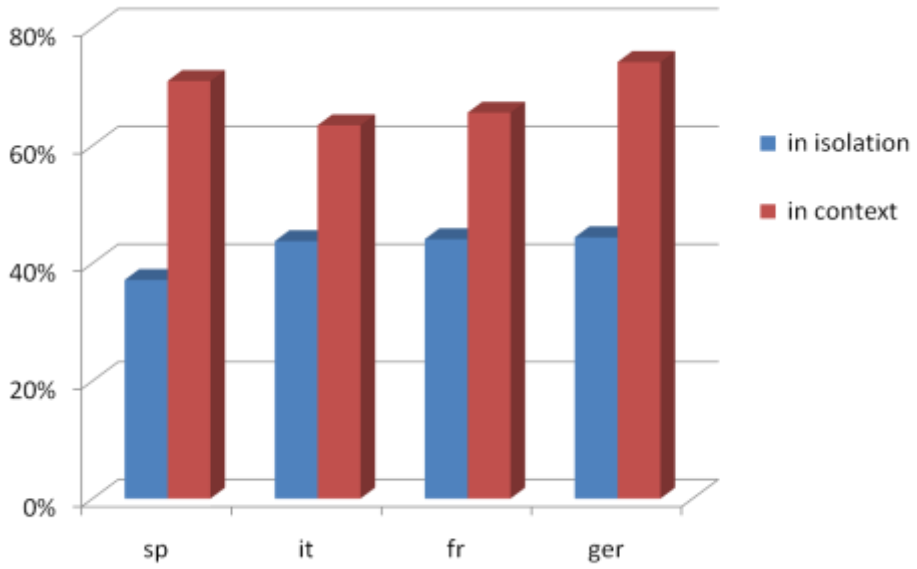


Figure 3: Percentages of correctly recognised words in isolation and context according to listeners' L1s.

5. Discussion

The analysis above shows that the results regarding aspiration and realisations of the interdental fricative conform to Jenkins' (2000) observations. There was a strong tendency that plosives with a longer VOT (40-60ms) were recognised more easily than those with a shorter VOT (0-20ms). Regarding the realisations of the phonemes [θ]/[ð], the realisations /s z/ gained lower scores for both the lenis and the fortis variant. Additionally, there was a tendency in the L1-German and L1-French groups towards displaying a slight advantage in recognising the interdental fricative when realised as an alveolar fricative. This might indicate an advantage of the familiarity with commonly used substitutions of the interdental fricative used by EFL-learners of German and French (who commonly substitute the interdental fricative with an alveolar fricative).

Regarding different realisations of /r/, the results differ from Jenkins' assumption that rhotic pronunciation aids intelligibility more than non-rhotic. All listener groups recognised the words which contained a non-rhotic realisation of /r/ more often than those which contained a rhotic pronunciation of /r/. Additionally, it needs to be mentioned that all words with non-rhotic realisation were pronounced with standard pronunciation. Therefore, it is not clear if the results point towards a high intelligibility of non-rhotic /r/, as standard pronunciation was generally highly intelligible. Moreover, the

considerable difference in preference for non-rhotic pronunciation was, apart from for the German-speaking group, only visible for the correct recognition of the words, not the feature. Moreover, there was a tendency showing that the uvular realisation of /r/ reached lower intelligibility rates with all listener groups but the French. This could be an indication of the fact that uvular /r/ is less intelligible than other realisations of /r/. These results regarding rhotic and non-rhotic variants as well as different places of articulation of /r/ add to Jenkins' findings and point towards an interesting tendency worth further investigation.

Regarding the relevance of listener- and item-related aspects, the results indicate that both could possibly influence intelligibility. However, the effect of listener-related aspects seemed to be rather inconsistent. For listener-related aspects, such as previous language experience, no correlation between overall familiarity with other accents, which listeners had indicated in the questionnaires, and intelligibility scores was found. The German-speaking group, who were not exchange students, had stated less experience with other accents but this had no visible effect on the intelligibility scores (cf. Table 6). However, the familiarity of accents could not be measured objectively and was subject to the participants' own estimations.

	Spanish	Italian	French	German
Average % of recognised words in total	14.1	16.6	16.8	16.9
Average rating of accent familiarity	8.0	10.8	7.9	4.9

Table 6. Comparison of correctly recognised words on average and average of indicated familiarity with accents according to listener groups.

For the correctly identified features in the words, both an effect of item-related as well as of listener-related aspects was observed. All listener groups had low intelligibility scores for the interdental fricative produced as /s z/ and, apart from the French, the uvular realisation of /r/. This low decoding of particular realisations, regardless of the listeners' L1, points towards the importance of item-related factors for intelligibility.

It would appear that familiarity with particular substitutions of the foreign languages sometimes, but not consistently, eases intelligibility. In the case of uvular /r/, which is common in French and German, the French-speaking group apparently benefited from familiarity with this realisation. However, for the German-speaking group this was not advantageous. These results are

in line with the studies of Smith & Bisazza (1982), Major et al. (2002) and Munro, Derwing & Morton (2006), in which it was observed that listening to one's own accent was not consistently beneficial for all tested listener groups.

As to the effect of phrasal and syntactic co-text on intelligibility, the results showed a considerably higher intelligibility for words in their co-text than in isolation. This implies that the listeners can benefit from the syntactic context given. This is not in line with e.g. Bond, Moore & Gable (1996) and Jenkins (2000) who propose that L2 listeners are primarily dependent on the acoustic signal and do not benefit much from knowledge of the context. At least in a phrasal and syntactic co-text, this cannot be confirmed by this study as the listeners often adapted their interpretations of the stimulus-words to the environment. This indicates that the co-text might play a major role for NNL in the interpretation of utterances.

Finally, it was found that standard pronunciation was relatively well understood for all three tested aspects, i.e. plosives with common VOT-length for English, the interdental fricative realised as [ð]/[θ] and /r/ realised as alveolar or retroflex approximant. It is important to add that most other realisations (apart from the ones mentioned) did not impair intelligibility to a large degree. However, the experience of the listeners with NS-language and context might have shaped these results. Whilst considering the limitations of this dataset, two suggestions may be made for the teaching of English phonetics and phonology for an ELF-speaking target group: Firstly, for the aspects tested in this study, the standard pronunciation, as it has been largely taught, is a variety of English which is intelligible to speakers of different first languages. Secondly, other variants of [ð]/[θ] and /r/ can be tolerated and possibly even taught, especially if this eases teachability and learnablity.

6. Conclusion

Although no definite conclusions can be drawn, due to the framework of this study with its aforementioned limitations, some clear tendencies have been illustrated. Firstly, Jenkins' observations about two of the three features, namely aspiration and different realisations of the interdental fricative, could be confirmed with regards to mutual intelligibility in ELF when tested with the dictation method. For the third aspect, realisations of /r/, some tendencies could be shown which might be interesting for future research. In order to be able to draw further conclusions for ELF-research and language teaching, the research would need to be extended to a larger dataset, a larger number of participants with a greater variety of first languages and elements on the segmental and suprasegmental levels, which were not considered in this

study. Additionally, a comprehensive investigation according to sub-categorisation into syllable positions of the phonemes and for realisations of /r/ according to manner of articulation would be necessary. This could also be extended to investigate the role of aspects such as vowel quality and quantity.

As ELF plays an increasingly important role around the world, there is a need for a greater number of studies on intelligibility in a NNS-NNL context. These will also largely contribute to answering other psycholinguistic questions, such as the differences in L1 and L2 speech perception, e.g. the degree to which top-down and bottom-up processes are involved in both.

Appendix

Text

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Separation of text for test 2

1) please call Stella.
2) ask her to bring these things with her
4) six spoons of fresh snow peas
5) five thick slabs of blue cheese
6) and maybe a snack for her brother Bob
7) we also need a small plastic snake
8) and a big toy frog
8a) for the kids
9) she can scoop these things
10) into three red bags
11) and we will go meet her
12) Wednesday at the train station

References

Anderson-Hsieh, Janet; Johnson, Ruth; Koehler, Kenneth. 1992. "The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure". *Language Learning* 42, 529-555.

Bent, Tessa; Bradlow, Ann. 2003. "The interlanguage speech intelligibility benefit". *Journal of the Acoustical Society of America* 114, 1600-1610.

Bond, Z. S.; Moore, Thomas J.; Gable, Beverly. 1996. "Listening in a second language". *Proceedings of the forth international conference on spoken language* 4, 2510-2513. www.asel.udel.edu/icslp/cdrom/vol4/038/a038.pdf (26 April 2007).

Bradlow, Ann R.; Pisoni, David M. 1999. "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors". *Journal of the Acoustical Society of America* 106, 2074-2085.

Brodkey, Dean. 1972. "Dictation as a measure of mutual intelligibility: A pilot study". *Language Learning* 22, 203-220.

Burda, Angela N.; Scherz Julie A.; Hagemann, Carlin F.; Edwards, Harold T. 2003. "Age and understanding of speakers with Spanish or Taiwanese accents". *Perceptual and Motor Skills* 97, 11-20.

Carter, Ronald. 1998. "Orders of reality: CANCODE, communication, and culture". *ELT Journal* 52, 43-56.

Cho, Taehong; Ladefoged, Peter. 1999. "Variations and universals in VOT: evidence from 18 languages". *Journal of Phonetics* 27, 207-229.

Crystal, David. 2000. "The future of English". In Lynch, Donal; Pilbeam, Adrian (eds.). *Heritage and Progress. Proceedings of the SIETAR Europa Congress 1998*. Bath: LTS Training and Consulting, 6-16.

Derwing, Tracey M.; Munro, Murray J. 1997. "Accent, intelligibility and comprehensibility – evidence from four L1s". *Studies in Second Language Acquisition* 19, 1-16.

Docherty, Gerald J. 1992. *The timing of voicing in British English obstruents. (Netherlands Phonetic Archives).* Berlin/New York: Foris Publications.

Einstein, Miriam; Verdi, Gail. 1985. "The intelligibility of social dialects for working class adult learners of English". *Language Learning* 35, 287-298.

Fayer, Joan M.; Krasinksy, Emily. 1987. "Native and non-native judgments of intelligibility and irritation". *Language Learning* 37, 313-326.

Field, John. 2005. "Intelligibility and the listener: The role of lexical stress". *TESOL quarterly* 39, 399-423.

Firth, Alan. 1996. "The discursive accomplishment of normality. On 'lingua franca' English and conversation analysis". *Journal of Pragmatics* 26, 237-259.

Flege, James E. 1988. "Factors affecting degree of perceived foreign accent in English sentences". *Journal of the Acoustical Society of America* 84, 70–79.

Florentine, Mary. 1985. "Non-native listeners' perception of American English in noise". *Proceedings of Inter-noise* 85, 1021-1024.

Foulkes, Paul; Docherty, Gerry. 2001. "Variation and change in British English /r/". In Van de Velde, Hans; van Hout, Roeland (eds.). *'r-atics: Sociolinguistic, phonetic and phonological characteristics of /r/. (Etudes & Travaux 4).* Brussels: Free University of Brussels, 27-44.

Hazan, Valerie L.; Markham, Duncan. 2004. "Acoustic-phonetic correlates of talker intelligibility for adults and children". *Journal of the Acoustical Society of America* 116, 3108-3118. http://scitation.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf&id=JASMAN000116000005003108000001&idtype=cvips&prog=normal (18 May 2007).

Hirschfeld, Ursula. 1994. *Untersuchungen zur phonetischen Verständlichkeit Deutschlernender.* Frankfurt am Main: Forum Phoneticum.

Jenkins, Jennifer. 2000. *The Phonology of English as an international language: new models, new norms, new goals.* Oxford: Oxford University Press.

Jenkins, Jennifer. 2002. "A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language". *Applied Linguistics* 23, 83-103.

Jenkins, Jennifer. 2003. *World Englishes. A research book for students.* London: Routledge.

Khattab, Ghada. 2000. "VOT Production in English and Arabic bilingual and monolingual children". *Leeds working papers in linguistics* 8, 95-122. http://www.leeds.ac.uk/linguistics/WPL/WP2000/Khattab.pdf (25 May 2009).

Ladefoged, Peter; Maddieson, Ian. 1996. *The sounds of the world's languages.* Oxford: Blackwell.

Ladefoged, Peter. 2005. *Vowels and consonants: an introduction to the sounds of languages*. (2nd edition). Malden: Blackwell.

Magen, Harriet. S. 1998. "The perception of foreign-accented speech". *Journal of Phonetics* 26, 381-400.

Major, Roy. C., Fitzmaurice, Susan F.; Bunta, Ferenc; Balasubramanian, Chandricka. 2002. "The effects of nonnative accents on listening comprehension: implications for ESL assessment". *TESOL Quarterly* 36, 173-190.

Meierkord, C. 1996. *Englisch als Medium der interkulturellen Kommunikation. Untersuchungen zum non-native/non-native speaker Diskurs.* Frankfurt am Main: Lang.

Munro, Murray J. 1998. "The effects of noise on the intelligibility of foreign-accented Speech". *Studies in Second Language Acquisition* 20, 139-154.

Munro, Murray J.; Derwing, Tracey M. 1995a. "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners". *Language Learning* 45, 73-97.

Munro Murray J.; Derwing, Tracey M. 1995b. "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech". *Language and Speech* 38, 289-306.

Munro, Murray J.; Derwing Tracey M.; Morton, Susan L. 2006. "The mutual intelligibility of L2 speech". *Studies in Second Language Acquisition* 28, 111-131.

Niedzielski, Nancy A.; Preston, Dennis R. 2000. *Folk linguistics.* Berlin [u.a.]: Mouton de Gruyter.

Osimk, Ruth. 2007. *Aspiration, [θ]/[ð] und /r/ in Englisch als Lingua Franca – eine psycholinguistische Studie zu drei Vorschlägen des Lingua Franca Core.* Unpublished MA thesis, University of Vienna.

Rajadurai, Joanne. 2007. "Intelligiblity studies: a consideration of empirical and ideological issues". *World Englishes* 26, 87-98.

Seidlhofer, Barbara. 2001. "Closing a conceptual gap: the case for a description of English as a lingua franca". *International Journal of Applied Linguistics* 11, 133-158.

Seidlhofer, Barbara. 2004. "Research perspectives of teaching English as a Lingua Franca". *Annual Review of Applied Linguistics* 24, 209-239.

Smith, Larry E.; Bisazza, John A. 1982. “The comprehensibility of three varieties of English for college students in seven countries”. *Language Learning* 32, 259-269.

van Wijngaarden, Sander J., Steeneken, Herman J. M.; Houtgast, Tammo. 2002. “Quantifying the intelligibility of speech in noise for non-native listeners”. *Journal of the Acoustical Society of America* 111, 1906–1916.

Widdowson, Henry G. 2004. *Text, Context, Pretext: Critical Issues in Discourse Analysis.* Oxford: Blackwell Publishing.

Yao, Yao. 2007. “Closure duration and VOT of word-initial voiceless plosives in English in spontaneous speech“. *UC Berkeley Phonology Lab Annual Report*, 183-225.

Zielinski, B. 2004. “Measurement of speech intelligibility: What are we actually measuring?” *Paper presented at the annual meeting of the American Association for Applied Linguistics,* Portland, OR.

Online Resources:

Weinberger, Steven. H. “Speech Accent Archive” http://accent.gmu.edu/ (2. January 2007).

How to contact us:

c/o
Institut für Anglistik & Amerikanistik der Universität Wien
Universitätscampus AAKH, Spitalgasse 2-4, Hof 8 .3
A – 1090 Vienna; Austria

fax	**(intern.) 43 1 4277 9424**
eMail	**theresa.illes@univie.ac.at**
	marie-luise.pitzl@univie.ac.at
W3	http://anglistik.univie.ac.at/views/
	(all issues available online)

IMPRESSUM:
EIGENTÜMER, HERAUSGEBER & VERLEGER: VIEWS**,** C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, A - 1090 WIEN, AUSTRIA**. FÜR DEN INHALT VERANTWORTLICH**: THERESA-SUSANNA ILLES, MARIE-LUISE PITZL **WEBMASTER**: MONIKA WITTMANN **REDAKTION**: HEIKE BÖHRINGER, ANGELIKA BREITENEDER, CHRISTIANE DALTON-PUFFER, OLGA FISCHER, CORNELIA HÜLMBAUER, JULIA HÜTTNER, THERESA-SUSANNA ILLES, BRYAN JENNER, GUNTHER KALTENBÖCK, THERESA KLIMPFINGER, URSULA LUTZKY, BARBARA MEHLMAUER-LARCHER, MARIE-LUISE PITZL, ANGELIKA RIEDER-BÜNEMANN, NIKOLAUS RITT, HERBERT SCHENDL, BARBARA SCHIFTNER, BARBARA SEIDLHOFER, UTE SMIT, LOTTE SOMMERER, BARBARA SOUKUP, JOHANN UNGER, H.G. WIDDOWSON. ALLE: C/O INSTITUT FÜR ANGLISTIK & AMERIKANISTIK DER UNIVERSITÄT WIEN, UNIVERSITÄTSCAMPUS AAKH, SPITALGASSE 2, A - 1090 WIEN. **HERSTELLUNG:** VIEWS