The Development of Specifications for Item Development and Classification within The Common European Framework of Reference for Languages: Learning, Teaching, Assessment

Reading and Listening

Final Report of

The Dutch CEF Construct Project

J Charles Alderson, Neus Figueras, Henk Kuijper, Guenter Nold, Sauli Takala, Claire Tardieu

July 2004

Contents

***** Available on request from the Project Coordinator, J Charles Alderson (c.alderson@lancaster.ac.uk)**

Executive Summary

The Common European Framework (CEF) is intended as a reference document for curriculum and syllabus development, textbook writing, teacher training, and for assessment. However, the CEF in its present form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEF. The Project described in this Report therefore addressed the following research questions:

- Do we have in the CEF an instrument to help us construct reading and listening items and tests based on the CEF?

- If the CEF scales, together with the detailed description of language use contained in the document, are not sufficient, what is needed to develop such an instrument, and what should the document be like?

The methodology of the Project involved gathering expert judgments on the usability of the CEF for test construction, identifying what might be missing from the CEF, developing a frame for analysis of tests and specifications, and then examining a range of test specifications, guidelines to item writers, and sample test tasks at the six levels of the CEF. Outcomes included a critical review of the CEF, a set of compilations of CEF scales and of test specifications at the different levels of the CEF, and a series of frameworks or classification systems, which led to a World-wide-web-mounted instrument, based on the CEF, to enable the characterisation of tests and items in relation to the CEF. The resulting Grid is available at www.ling.lancs.ac.uk/cefgrid.

Analyses of the application of earlier versions of the Grid to sample test items and texts showed that inter-analyst agreement was quite promising, but needs to be improved by training and discussion before decisions are made. The relation between the dimensions in the Grid, and individual CEF levels was not (yet) very obvious. Relatively few dimensions showed any significant association with the six CEF levels. However, the collection of much more extensive data using the Grid is recommended before solid conclusions can be reached about the relationship, or lack of it, between the dimensions of the Grid and CEF levels. It is concluded that the Grid is a useful instrument for the description of test items and tasks in terms of the CEF. A series of recommendations are made for revision and development of the Grid and associated guidance and training materials, and for further research using the revised Grid.

The Report concludes that the identification of separate levels in the CEF is at least as much an empirical matter as it is a question of the content of the tests as determined by test specifications or as identified by any content classification system or Grid.

The Project team therefore proposes a set of procedures, which in essence involve analysing the content of the test in question using the Grid, estimating the CEF level of items, texts and tests, and then pre-testing and calibrating the items, as well as conducting standard-setting procedures to set the boundaries of the CEF levels on the scale coming from the calibration, assigning a psychometric level to the items, and then assigning a definitive CEF level if the psychometric level falls within the band of the estimated level (in other words if the estimation based on the **analysed content** is comparable with the **psychometric** value).

Introduction

This report describes the work of a Project funded by the Dutch Ministry of Education, Culture and Science and accordingly dubbed the Dutch CEF Construct Project, whose purpose was to develop an instrument, based on the Council of Europe's Common European Framework of Reference (CEF) as far as possible, that would describe the construct of reading and listening, for English, French and German, that should underlie test items, tasks and whole tests at the six levels of the CEF. Such an instrument is intended to provide guidance to item writers on how to construct new items or test tasks and how to analyse existing test items at the various CEF levels, as well as guidance to item bank builders on the design of an item bank based upon the CEF, and on how to select items for inclusion in such an item bank.

The Common European Framework is intended as a reference document for curriculum and syllabus development, textbook writing, teacher training, and for assessment. The CEF contains a rather comprehensive review of the elements that play a role in the teaching and learning of languages. But it also includes a number of scales which describe a series of levels of language proficiency and which have received considerable attention from professionals. The CEF is increasingly referred to across Europe, and there is therefore an urgent need to illustrate the levels of the CEF with calibrated test items. It is hoped that eventually it will prove possible to construct an item bank that can serve as a common operational tool that would enable the linking of national tests and examinations to the CEF.

The experience of a number of previous projects, including the EU-funded DIALANG Project, suggested that the CEF in its present form may not provide sufficient theoretical and practical guidance to enable test specifications to be drawn up for each level of the CEF. Whereas the illustrative CEF scales for the productive skills appear to be quite adequate for the assessment of written and especially spoken performance, in the case of the receptive skills the empirical evidence to justify the scales that exist is not as strong and thus the wording of the descriptors for receptive skills is unlikely to be sufficiently explicit and precise for test specifications to be developed. Thus the Project team expected that it would have to do some further work to make this possible. Such adaptation is envisaged and endorsed in the CEF itself, as will appear from the following brief review of the CEF approach.

In order to fulfil its functions, the Common European Framework (CEF) was planned to be comprehensive, transparent and coherent.

"By 'comprehensive' is meant that the Common European Framework should attempt to specify as full a range of language knowledge, skills and use as possible (without of course attempting to forecast *a priori* all possible uses of language in all situations – an impossible task), and that all users should be able to describe their objectives, etc., by reference to it. The CEF should differentiate the various dimensions in which language proficiency is described, and provide a series of reference points (levels or steps) by which progress in learning can be calibrated. It should be borne in mind that the development of communicative proficiency involves other dimensions than the strictly linguistic (e.g. sociocultural awareness, imaginative experience, affective relations, learning to learn, etc.).

By ‘transparent’ is meant that information must be clearly formulated and explicit, available and readily comprehensible to users.

By ‘coherent’ is meant that the description is free from internal contradictions.” (CEF, 2001, page 7)

It is obvious that these are very ambitious goals and their implementation may therefore have been only partial.

The CEF emphasises that the construction of a comprehensive, transparent and coherent framework for language learning and teaching does not imply the imposition of one single uniform system.

“On the contrary, the framework should be open and flexible, so that it can be applied, with such adaptations as prove necessary, to particular situations. The CEF should be:
• *multi-purpose*: usable for the full variety of purposes involved in the planning and provision of facilities for language learning
• *flexible*: adaptable for use in different circumstances
• *open*: capable of further extension and refinement
• *dynamic*: in continuous evolution in response to experience in its use” (CEF, 2001, page 7-8)

The Dutch CEF Construct Project was carried out following this basic CEF philosophy. It accepted the challenge of applying the CEF to a special situation and making adaptations that might prove necessary. A key decision in the Project was to exploit the CEF as much as possible, identify gaps and areas that needed clarification and produce a document which would serve the Project goals. However, the Project team also strongly believes that the CEF itself needs and deserves to be developed to make it an even better tool for language education. The Project team feels that it is not only the right but also the duty of the users of the CEF to try to make contributions to this development work. In this way, the CEF can serve properly as a common and European framework. The Dutch CEF Construct Project hopes it has provided a concrete contribution to such development work.

Research questions and tasks

The basic questions the Project asked were:

- Do we have in the CEF an instrument to help us construct reading and listening items and tests based on the CEF?

- If the CEF scales, together with the detailed description of language use contained in the document, are not sufficient, what is needed to develop such an instrument, and what should the document be like?

After detailed inspection of the extent to which the CEF itself serves as a basis for test specifications and how it needs to be complemented and modified so as to eliminate ambiguities, the Project planned to:

- develop a Frame of Analysis of items and tests of reading and listening in English, French and German
- examine a range of items and tests claimed to be at the various levels of the CEF
- examine what the tests have in common in their test specifications and how they differ
- examine how the tests operationalise in test tasks the development of language ability

From such an investigation we hoped to develop a more specified theoretical framework and a practical instrument to complement the CEF itself. Hopefully the insights gained will lead to the development of guidance to test developers on how to analyse and construct items and tests of reading and listening at the various levels of the CEF.

Issue to be addressed

The CEF, being such a comprehensive description of language use, can also be considered, implicitly at least, as a theory of language development. However, the Can-Do scales for reading and listening present a taxonomy of behaviours rather than a theory of development in listening and reading abilities. Moreover, it is far from clear that the still relatively abstract Can-Do descriptors in the CEF can be turned into items that illustrate or exemplify the different CEF levels. The experience of the DIALANG Project was that it was necessary to develop additional specifications before the CEF could be used as the basis for test development. DIALANG is, however, only one example, and is *sui generis* because it has developed diagnostic tests for delivery by computer across the Internet. It could not be assumed that the DIALANG experience and specifications would generalise across the variety of assessment contexts in Europe. On the other hand, the DIALANG documents were considered to be a very useful synopsis of the various scales in the CEF, enabling test writers to keep the rich variety of descriptions in mind.

It is essential that the research questions above be answered before attempts are made to link tests and examinations to the CEF levels or before any potential European item bank for reading and listening is developed. Unless we have such an instrument, we will not, for example, be in a position to select suitable items for inclusion in an item bank on a principled – i.e. theoretical – basis rather than simply on psychometric criteria. What is needed is an instrument that contains test-relevant linguistic, psycho-and socio-linguistic as well as pragmatic criteria for text and task selection at different CEF levels and for item construction or revision.

In what follows, and basing ourselves on the CEF (2001 English version, page 9) we can define a reading or listening item as

"*A task (which has an intention/ goal/ purpose) for an individual or group, requiring an understanding of a text related to a theme in a domain which requires certain strategies, under certain conditions and limitations*". This definition presents a number of problems for item bank and test constructors, as we shall see in due course.

Method

Essentially the methodology of the Project involved gathering expert judgments on the CEF to identify what might be missing from the Framework, to develop a frame for analysis of tests and specifications, and then to examine a range of test specifications, guidelines to item writers, and actual sample test tasks at the six levels of the CEF. The outcome was expected to be a critical review of the CEF and an instrument that would complement the CEF itself.

Although solid theoretical foundations may be lacking, it is clearly the case that there is a great deal of experience across Europe in producing tests and examinations at a range of different levels of ability, and since many of these tests are explicitly claimed to be at the various levels of the CEF, it made sense to examine these tests and examinations to see what they had in common, how they differed, and how they operationalised in test items and tasks the development of language ability. From such an investigation it should be possible to develop a more specified draft theoretical framework and a practical instrument to complement the CEF itself.

A group of six language testing experts was convened, representative of a range of different testing and assessment cultures across Europe, in order to identify potentially relevant documents, and to examine them for insights that could lead to the construction of a set of guidelines for test developers on how to construct both items and tests at the various levels of the CEF. These experts have wide theoretical knowledge and practical experience in test construction, as well as familiarity with the diversity of assessment contexts in Europe, and with use of the CEF in language education generally.

The method used by the Project team was inductive in the sense that the results obtained in each stage of the procedure were used to reflect on the outcomes, to plan the next stages and to revise and extend the analytical tools as more experience was accumulated. The strategy in developing the analytical tools was first to adhere to the exact wording of the CEF and then to make adaptations as they were considered relevant. Continuous discussion, both focused and spontaneous, using email and during meetings, was a crucial element of the Project methodology.

The key products of the Project are the grids of reading and listening items developed through several versions by applying them in analysing and judging items (Appendix 1). The ratings were done independently by the experts, which provided a good test of their feasibility and demonstrated the challenge in reaching an adequate level of agreement.

These grids were revised five times on the basis of the results of the analyses of analyst agreement, on what specific problems in using the grids the analysts reported and what suggestions they made for how the grids could be improved. The table below presents the activities and drafts developed during the Project, from analysing the CEF to using a recent version of the Grid to analyse reading and listening items and test specifications. The items in bold are contained in Appendix 1 to this Report. A detailed timetable is in Appendix 2.

Draft Frame of Analysis for Listening (by CEF level) (**CEF Frame 1**)
Draft Frame of Analysis for Reading (by CEF level) (**CEF Frame 1**)

Compilation of CEF scales for Reading and Listening, by level. (**Appendices 3 and 4**)
Compilation of extracts from the CEF for Reading and Listening (**Appendix 5**)
Compilation of extracts from preliminary Manual for Reading and Listening (**Appendix 6**)

Application of CEF Frame 1 to test items in order to assess applicability
Critical inspection and revision of draft Frames of analysis.

Revised Frame 2 for Reading (**CEF Frame 2**)
re-Revised Frame 2 for Listening (**CEF Frame 2**)
Grid for analysing Listening tasks (**Grid 1**)

Revision of Frames into Grid, trialling on test items and further revisions

Reading Grid (revised) (**Grid 2**)
Listening Grid (revised) (**Grid 2**)
Grids and Guides for analysis of reading and listening (6 Guides) (**Grid 2: Guides**)

Analysis of DIALANG reading and listening items using Grid 2 and Guides.

Revision of Grid 2 and incorporation of Guides into a Web-based template (**Grid 3**)

Analysis of anonymous Reading tasks using Grid 3 by five analysts

Analysis, using Grid 3, of available test specifications for Finnish National Certificates in 9 languages, Catalan exams for 12 languages, Cambridge ESOL Main Suite exams, Profile Tests Dutch as a second language, Certificats de français and Zertifikat Deutsch exams.

Correction and discussion of data input via Grid 3 to database
Discussion of results and proposals for revision of test specifications analysis
Analysis of tests produced in France, using Grid 3
Analysis of Catalan tests of French and German using Grid 3

Compilation of results of specifications analyses
Revision and submission of analyses of test specifications
Revision of Grid 3 in light of analysts' recommendations (**Grid 4**)
Field testing and further development of Grid 4

The activities above can be divided into three phases, which are described below.

Phase One: Analysing the CEF

Initially, the Project team examined the CEF itself to identify precisely where the gaps were, if any. The Project team produced a compilation, from the overall and illustrative scales for reading and listening from the CEF, of the Can-Do statements at each of the six levels (A1 to C2) of the CEF (Appendices 3 and 4). The team also compiled a selection of extracts from the CEF that had any conceivable relationship to reading and listening abilities (Appendix 5). In addition the Project produced a compilation of extracts from the preliminary Manual for linking exams to the CEF that were relevant for reading and listening tests, for further reference (Appendix 6).

The team then sought to extract from these scales, and associated CEF text, features relevant to test design that the scales contained at any given level. For example, all the Can-Do statements begin with a verb which appears to characterise aspects of the nature of comprehension (*understand, recognise, locate, infer* and so on). Such features were termed "Operations" and a category of Operations constituted the first column in a draft frame of analysis intended to characterise what the CEF says about comprehension at each level. Similarly, the Can-Do statements describe "What" somebody can comprehend at any given level, often in terms of the meaning of a text, or the language of the text, and so on. The Source Texts that learners are said to be able to comprehend at any given level constitute a third column in the instrument, and so on.

As a result of this work, the Project developed a draft Frame of Analysis, based on the CEF, which it critiqued for internal consistency and comprehensibility. It then sought to apply the Revised Draft Frame to a number of sample tests of reading and listening. During this process it rapidly became clear what were the major gaps in the CEF (and thus in the revised Frame), as well as what needed further and more consistent definition before it could usefully be included in any instrument intended to characterise test tasks or items at any given CEF level. An account of these gaps, missing definitions and ambiguities will be presented in the next section, dealing with the results of Phase One.

Phase Two: Applying the Frame of Analysis

The next phase of the Project involved revising CEF Frame 2, by adding elements not in the CEF which were thought to be essential when seeking to characterise tests of reading and listening. This new instrument, dubbed a Grid to distinguish it from the CEF-based Frames, was then developed into a second version, Grid 2, which included elements that need to appear in any instrument that is intended to characterise items in an item bank for reading and listening.

It is clear that such an instrument is beginning to resemble the sort of contents that are required in any language test specification. Indeed, one possible way of working would have been to start with a statement of what should be contained in any language test specification and then to search the CEF for such elements. However, the Project team deliberately chose not to do this in the initial stages of the Project, partly because such a procedure risked imposing on the CEF an alien framework. It was thought likely to be the case that test specifications would vary across Europe. Most textbooks on specifications for language tests have a North American or English-speaking origin, rather than a European, multilingual source (see, for example, Alderson et al, 1995,

Bachman and Palmer, 1996, and Davidson and Lynch, 2002). Instead, the team's work was deliberately grounded in the CEF, working as close to the scales as possible, and only clarifying and adding to the Frames and Grids when the team was certain that the CEF itself did not provide the necessary information.

Grid 2 was then applied to a set of reading and listening items from DIALANG and, in the light of the analysis of the results (see Appendix 9) the Grid was developed further, into a Web-based tool labelled Grid 3.

Phase Three: Applying the revised frame of analysis to a wider range of test specifications and items

In the third phase of the Project, Grid 3 was used to analyse a range of test tasks and items from a variety of sources (see results in Appendix 11 and below). In addition, the Project team examined in detail the specifications and guidelines that a range of European testing and examination authorities use to write items and construct tests at the different levels of the CEF. Bodies that cooperated included DIALANG, members of the Association of Language Testers in Europe (ALTE) as well as other bodies with experience of writing tests at the CEF levels, like the Escuelas Oficiales de Idiomas in Catalonia, Spain, and national authorities responsible for school-leaving examinations, in so far as these have an established, or at least asserted, relationship to the six levels of the CEF. For a sample of the letters sent to exam providers, see Appendix 7, and for a list of materials received, see Appendix 8.

During the different phases of the Project, the team examined specimen tasks and specifications from DIALANG, the Dutch school-leaving examinations HAVO 2000 and MAVO 1999, The Profile Test Dutch as a Second Language, The Finnish Matriculation Examinations, The Finnish National Certificates for English, The Catalan Official Schools of Languages Examinations for English, French and German, the French Certificate of Higher Education in Foreign Languages, the Diploma of Language Competence, the Baccalaureate, Cambridge-ESOL's Certificates in English Language Skills, Cambridge ESOL's Main Suite of English exams, and the Certificats de français and Zertifikat Deutsch produced by WBT (Weiterbildungs-Testsysteme GmbH).

Results of Phase One

The first outcome of Phase One was a compilation of all the reading and listening scales in the CEF, level by level rather than the organisation activity by activity currently in the CEF (see Appendices 3 and 4). This has since proved extremely useful for familiarisation purposes, but importantly it enabled us to identify a number of problems with the CEF.

Problems with the CEF

The major problems were of four types:

1. Terminology problems: synonymy or not?
2. Gaps, where a concept or feature needed for test specification or construct definition is simply missing

3. Inconsistencies, where a feature might be mentioned at one level but not at another, where the same feature might occur at two different levels, or where at the same level a feature might be described differently in different scales.
4. Lack of definition, where terms might be given, but are not defined

These problems are illustrated below

Terminology problems: synonymy or not?

The CEF uses a variety of different verbs to indicate comprehension, some of which can stand alone, but others require a noun phrase.
For example A2 Listening uses

- Understand
- Take
- Get
- Follow
- Identify
- Infer

B1 for Reading uses 7 different verbs:

- Understand
- Locate
- Scan
- Identify
- Combine
- Extrapolate
- Recognise

whilst B2 uses 8 verbs:

- Understand
- Scan
- Monitor
- Obtain
- Select
- Evaluate
- Locate
- Identify

Often different words are being used synonymously, possibly for stylistic reasons or because the Can-Do statements were originally derived from a wide range of different taxonomies: thus

"*I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues*" (page 26) and

"*Can recognise familiar names, words and very basic phrases on simple notices in the most common everyday situations*" (p70).

Clearly *understand* and *recognise* are synonymous here, but it is unclear whether there is a meaningful distinction between the two main verbs in the following:

"*Can **identify** the main conclusions in clearly signalled argumentative texts. Can **recognise** the line of argument in the treatment of the issue presented, though not necessarily in detail*."

Are *find* and *locate* synonyms in the following:

*Can **find** specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables. Can **locate** specific information in lists and isolate the information required (e.g. use the 'Yellow Pages' to find a service or tradesman)?*

We decided to standardise the terminology and its consistent use as much as possible. Since there is no description in the CEF of how cognitive operations might differ at different levels (or even whether they do) there is no basis in the CEF itself for standardising or grouping verbs, and we had to have recourse to theories of comprehension to resolve this issue, in Phase Two.

Gaps

A feature was considered to be missing if it was mentioned in general terms somewhere in the CEF but then was not distinguished according to the six levels of the CEF, or was not even specified at one level.

The first major gap in the CEF we identified was a description of the operations that comprehension consists of and a theory of how comprehension develops.

Related to this is the absence of any specification of micro-skills or subskills of comprehension. The one most immediately noticed was "*skim*" but others like "*distinguish relevant from irrelevant details*" or "*discriminate between fact and opinion*" also seem to be absent.

The text of the CEF introduces many concepts which are not then incorporated in the scales or related to the six levels in any way. These include the following (page numbers refer to the 2001 English edition of the CEF):

competence, general competence, communicative language competence (pages 9, 13, 108 ff)
activities, processes, text, domain, strategy, task (pages 10, 14, 15, 16)
context (pages 48/9, Table 5)
ludic and aesthetic uses of language (pages 55-6)
texts (page 93 ff)
text to text activities (page 100)
socio-cultural knowledge (pages 102-3)
study skills (pages 107-8)
tasks, including description, performance (conditions, competences, linguistic factors), strategies, difficulty (pages 157-166)

One major element that is missing from the CEF is The Task: what is it that candidates have to do with text? Although a whole chapter of the CEF is devoted to this topic, at no point is there a discussion of how tasks might be distinguished by level. Some of the illustrative scales are subdivided by task in a sense, since they address things like:

- Listening as a member of a live audience
- Reading for orientation
- Reading for information and argument

But other illustrative scales address specific texts:

- Listening to announcements and instructions
- Listening to radio and audio recordings
- Reading instructions.

In short, there seems to be no principled way in which such illustrative scales have been drawn up, and the dimension of Purpose - why one is reading or listening to any given text in any particular setting - is not addresed systematically at all. This gap is a serious problem for test writers and item bank builders.

In tests, however, in a sense the test method IS the task, and so a consideration of test methods is crucial. For multiple-choice methods in particular, the nature of the options offered should be considered part of the text to be processed, but distinguished from the input text. How the options are constructed, what content they have, how they are worded, in what order they appear, how many pieces of information in a text they address: all these and more will necessarily add to the difficulty or ease of the item, but at present the CEF has no way of taking this into account because it focuses exclusively on "action" and "real-world" use. Discussion of test method is absent from the CEF. However, although item writers need to know what test method to use at which CEF level, it is likely that such method effects will generalise to more than one CEF level, and they are unlikely to be defining characteristics of listening or reading tasks / texts at one level and not another. Nevertheless, the processing demands they create need to be taken into account somehow when devising specifications and giving guidance on what level a "performance" is at.

Inconsistencies

Many formulations in the Can-Do statements are not consistent. Sometimes similar descriptions are found at different levels. Sometimes at one level (B1) something is said about vocabulary in texts, while at lower or higher levels nothing is said about vocabulary.

It is unclear why the operation *recognise* is only mentioned at levels A1, B1 and C1 and not at A2, B2 and C2. This cannot be a principled omission.

Despite the proliferation of verbs, there are, nevertheless, inconsistencies in the use of different verbs. *Infer*, for example, only appears at some levels and not others, yet *inferencing* may well be

needed even for A1 items. Certainly by B1 one would expect *infer* to appear as an operation, and therefore also at B2. Yet it only appears in C1.

The use of a dictionary is not mentioned in the CEF at the lower levels, but only at B2. At C1 we find *occasional use of dictionary* but not at C2. It is unlikely to be the case that one defining characteristic of a C2 person is not using a dictionary.

In Listening, there are particular inconsistences with type of speech:

- Clear, slow and carefully articulated speech (A1)
- Clear, slow and articulated speech (A2)
- Clear, standard speech, familiar accent (B1)
- Normal speed, standard language (B2)
- For C1 and C2 no limitations are set on speech

Speed is not mentioned at B1; *standard* is first mentioned at level B1 and not at levels A1 and A2.

A feature may appear in one descriptor for a level, but not in another for the same level. For example, what is the difference between *specific information* (A2), and *specific predictable information* (A2)?

There is a misleading inconsistency in the mention of specific text types at some levels in the CEF but not at other levels. For example, advertisements are said to be processable at A2: "*Can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus, reference lists and timetables*". The only other reference to advertisements is at B2, where accommodation advertisements are specifically mentioned: "*CAN understand detailed information, for example a wide range of culinary terms on a restaurant menu, and terms and abbreviations in accommodation advertisements*." But it cannot be the case that the ability to understand any advertisement (except for accommodation advertisements) is already developed at A2.

Similarly "*Simple notices*" can be read at A1, "*everyday notices*" at A2. There are no other references to "*notices*" but one can certainly envisage "*notices*", for example the regulations on permitted and forbidden activities in public parks, which are very hard to understand, albeit "*everyday*".

<u>Lack of definitions</u>

Many terms are used in the CEF but are undefined. For example, "*simple*" is frequently used in the scales, but it is not clear how one is to decide what is simple and what is less simple, and especially "very simple". The CEF is language–independent, and so does not contain any guidance, even at a general level, of what might be simple in terms of structures, lexis or any other linguistic level. Therefore, the CEF would need to be supplemented with lists of grammatical structures and lexical items for each language to be tested, or it could recommend the use of electronic corpora, which could be referred to if terms like "*simple*" or "*frequent*" are to have any meaning for item writers or item bank compilers. Of course, what is 'simple' for one first language background might be far from simple for somebody with a different first language,

and therefore some appeal must be made to second language acquisition (SLA) theory or research. This may prove to be an intractable problem for tests intended for multilingual audiences.

The same definitional problem applies to many expressions used in the CEF scales: for example "*the most common", "everyday", "familiar", "concrete", "predictable", "straightforward", "factual", "complex", "short", "long", "specialised", "highly colloquial"* and doubtless other expressions. These all need to be clarified, defined and exemplified if items and tasks are to be assigned to specific CEF levels.

However, what is familiar in one culture, with particular background knowledge and expectations, may not be at all familiar in other cultures (or individuals). How this can be taken into account by item writers is far from clear, even though it may make sense in a self-assessment scale, since individual respondents can decide for themselves what is familiar, everyday, specialised, and so on. Even so, individuals cannot decide for themselves what is "*short"* or "*long"*.

Frames of analysis

A third outcome of Phase One, in addition to the compilations and the identification of gaps in the CEF, was the development of frameworks of analysis, based on the CEF initially, but then extended once the gaps had been identified. A summary table showing the various stages of development of the Frames is contained in Appendix 2. Given that the Frames were based on the CEF, the same problems identified in the CEF were necessarily also contained in the Frames. In addition, the Project team found the categories of Conditions and Limitations to contain an unprincipled mix of different aspects of comprehension, and the category of strategies could not easily be distinguished by level. Since it was necessary to fill the gaps in the CEF and the Frames, a new instrument, which was now dubbed a Grid, to distinguish it from the Frames which had been closely based on the CEF, was proposed at the end of Phase One. Grid 1 added dimensions which appeared to be needed, even though they do not appear in the CEF.

The classification of texts was thus subdivided into the following dimensions:

- Text type (factual, persuasive, argumentative)
- Text source (radio, daily life, public announcement)
- Text length
- Text 'interaction' (monologue, dialogue, multiple speakers)
- Text difficulty characteristics. This dimension could be operationalised by aspects from the CEF like:
  - Speed (listening)
  - The standard-non standard dimension (listening)
  - Vocabulary (from very simple to complex)
  - Pronunciation (from clear and fully articulated to casual speech) (listening)

The Project team also added the dimension of item type. See Grid 1 in Appendix 1.

An empirical process

The final major outcome of Phase One emerged from discussions about the role of test piloting, calibration, and empirical information on task difficulty. The basic conceptual problem faced was: what does it mean for an item to be at a level? People are said to be at B1 if they can do the things described at that level (to a satisfactory degree), but not (yet) most of the things one should be able to do at the next higher level. But the problem this Project was intended to address was what is the construct validity of concepts like 'B1'? Any test developer has to show that they can build measuring instruments that can classify people truly at the level to which they belong. However, this presents a circular problem: to validate the theory, measurement instruments are needed, but to validate these measurement instruments a theory is necessary on which one can rely.

Moreover, in order to know whether a given item is indeed at the level of difficulty intended, piloting on suitable samples of test-takers is crucial. But to do this, a suitable sample is needed, i.e., it is necessary to know the level of the test takers in order to judge the adequacy of the item. A problem of circularity therefore presents itself.

In order to escape from this circle, the Project team therefore proposed a set of procedures, as follows:

- Describe the text and items using the dimensions of a classification system (The Frames and Grids).
- Make a guess at the level of an item (guided by the classification system and the CEF scales), leading to an estimated CEF-level.
- Pre-test the items thus labelled, describing in detail the characteristics of the pilot sample.
- Calibrate the items.
- Do standard setting to set the boundaries of the levels on the scale coming from the calibration.
- Assign a psychometric level to the items.
- Assign a definitive level to the items. An item can only be assigned to a definitive level if the psychometric level falls within the band of the estimated level (in other words if the estimation based on the **analysed content** is comparable with the **psychometric** value).

This proposed process will be referred to again in the conclusions to this Report.

Results of Phase Two

During Phase Two, the Project team developed the frameworks of analysis, now called Grids, in face-to-face and email discussions, making reference to theories of comprehension and of assessment, as well as by referring to the experience of Project members. Grid 2 was in effect a set of open categories, but it was accompanied by a set of Guides as to how to complete the Grid, dimension by dimension (see Appendix 1). Once we had developed Grid 2, the first outcome of Phase Two, we then used it to categorise those DIALANG Reading and Listening items which had been submitted to the Council of Europe for inclusion in their reference materials for The

Manual. The aim of this analysis was two-fold: to test the Grid, to identify difficulties in using it, the ambiguity or indeterminacy of categories, and so on; and to see what sort of agreement could be reached by analysts who were very familiar with the CEF. The detailed analysis of the results of using Grid 2 is contained in Appendix 9, and summarised below.

Result of using Grid 2

Two sorts of dimensions appear in Grid 2. One sort can be characterised objectively - number of words, readability of text, length of a spoken text in time, and so on. The second sort can only be characterised by subjective judgements on the part of analysts. If we look at the second set of dimensions first, the amount of agreement among analysts varied quite widely across dimensions in the Reading and Listening Grids, but also within dimensions across items. Average agreement above 75% was only reached on Text Source and Topic, for Reading, and Accent, Topic and Operation, for Listening. However, several variables did not result in much discrimination across the levels in any case. This was particularly true for the Listening dimensions of Speed, Accent and Pronunciation, but also to some extent for Grammar and Vocabulary in both Reading and Listening. Despite this relative lack of discrimination, several of the non-discriminating dimensions resulted in average agreement of less than 60%, suggesting that we needed to find better ways of characterising items. Those dimensions that resulted in only 60% agreement or less were, for Reading, Text Structure, Domain, Grammar, Vocabulary, and What. For Listening, the dimensions were Speed, Domain, Grammar, Vocabulary, What, Discourse Type and Text Structure.

Thus, across Listening and Reading, only Topic resulted in good agreement, and particularly weak were Grammar, Vocabulary, Text Structure, Domain and What. When comparing the Operations tested, according to the Grid, with the subskills DIALANG claims are tested by the items, there was only 47% agreement (8/17) in the case of Reading, and 69% agreement (11/16) in the case of Listening.

However, it must be remembered that for many categories, analysts were using their own concepts and words to describe features, and although the analysis took account of what appeared to be similar underlying concepts, this freedom for individual analysts to identify what they considered to be critical features almost inevitably led to a degree of disagreement. One outcome of this analysis was agreement that further revisions of the Grid should seek to reduce this freedom to use one's own words as far as possible, by obliging analysts to select from restricted sets of features when characterising items and texts.

No analysed dimension on the Grids resulted in a significant association with the CEF level of the items studied. (DIALANG items had been constructed according to the CEF as far as possible, and had undergone empirical calibration and standard setting of the sort described in the Manual, thus their CEF levels were known.) In other words we had failed to find any variables which could adequately characterise the differences among CEF levels, and therefore we were not yet in a position to typify the difference between any two CEF levels, at least in terms of the dimensions identified in Grid 2. This meant that, in principle at least, any Text Source, Discourse Type, Text Structure, Domain, Topic, and so on, could appear at any CEF level. On this evidence, we were not yet in a position to conclude that, for example, certain Text Sources,

Discourse Types, etc do not appear at level A2, or tend to be confined to B2 or above, or whatever.

The objective dimensions resulted in only slightly less discouraging results. Length of written text in number of words correlated significantly with CEF level for Reading, but text readability (as measured by Microsoft Word) did not. No objective variable showed any association with CEF level, for Listening, including number of participants in the discourse, text length, or item type.

Comments from analysts on the experience of analysing items according to Grid 2 were wide-ranging. In general, the task of using the Grids was felt to have been useful but difficult, because of the number of dimensions, a lack of clarity as to the meaning of some dimensions, and the difficulty of assigning a level to an item, having characterised it along each dimension.

Dimensions considered not to have been much help in assigning a level to an item included Discourse Type, Text Structure, Domain and Topic. However, opinion was divided as to whether such dimensions should therefore be removed from the Grids. Although they did not help to identify a CEF level, it may nevertheless be useful for item writers to consider such dimensions in order to ensure better sampling of the construct, and where necessary to check if the test is representative, regardless of level. In fact, there was general agreement that two sorts of instruments are desirable: a Grid to help categorise items at CEF levels and Guidelines to Item Writers on what should be included in Reading and Listening tests, regardless of level. The latter would be comprehensive and would not distinguish features by level. It would thus not directly relate to the objective of this Project. The former might prove very difficult to construct.

With respect to the various dimensions of the Grid, it was felt that Operations needed clarification, possibly with a restricted number of verbs. It was generally agreed that What was a difficult dimension and needed further clarification. It was felt that Item Type and Text source should be kept, although neither seemed to relate to levels. Proposals for revisions of Grid 2 were made, and eventually resulted in Grid 3. Since the process of inputting one's individual analysis onto electronic forms, and then the Coordinator compiling results by cutting and pasting had proved very labour-intensive and time-consuming, it was agreed to explore the possibility of developing a Web-based Grid in Phase Three, with drop-down menus containing the features of a dimension wherever possible (which would in effect incorporate the previously separate Guides).

It was agreed that once the Grids had been revised, they should be applied to a wider range of items, since the computer-based DIALANG items contained only short texts, and typically had only one item per text, unlike the majority of paper-based test tasks the Project members were familiar with.

Results of Phase Three

In Phase Three, Grid 3 was developed, as described in detail in Appendix 10, and was used both to characterise a range of tasks from different sources, as had been recommended at the end of Phase Two, and also as a framework for the analysis of the test specifications we had received from exam bodies. The detailed analysis of the results for test tasks is contained in Appendix 11,

and the analysis of using the Grid as a framework for the analysis of specifications is contained in Appendix 12. The results of the analyses are summarised in the next two sections.

Use of Grid 3 to analyse test items and tasks.

It was considered important for all analysts to complete all analyses, as a further test of the transparency and applicability of Grid 3, and therefore all items selected had to be in English. Accordingly, the Project Coordinator selected tasks from the following sources, whose items had been empirically analysed and related to the CEF:

Cambridge ESOL: PET (=B1); FCE (=B2); CPE (=C2). Sample tasks in publicly available Handbooks
The Catalan Official Schools of Languages Exams: Elemental (=B1); Aptitud (=B2)
Finnish Matriculation Examinations: Mixed levels

Two testlets were selected from each level available, anonymised, placed in random order and then compiled into a booklet of 77 items, giving 16 tasks (Finnish 6, Catalan 4, Cambridge 6). Each analyst was asked to complete Grid 3 without discussing results with colleagues.

Average agreement of over 75% was achieved on the dimensions Authenticity, Domain and Broad Discourse Type, and Text Source came close at 73.75%. Agreement of less than 60% only occurred on Task Level Estimated. This was notably better than was achieved in Phase Two. The only dimension where agreement was less was on Topic (62.5% Phase Three, 77.6% Phase Two) but this was likely to be due to the fact that in Phase 2 the Coordinator had to interpret the wordings of the analysts, whereas in Phase Three analysts merely selected from a list. However, there were striking differences among analysts in terms of the frequency of use of the different dimensions in the Grid, especially for Item type, Text source, Topic, Vocabulary, Grammar, and Operation. Clearly, different analysts use the various categories differently. Individual input to a Grid will likely result in disagreement and discrepancy, and therefore it is essential that Grid users receive familiarisation and training in the use of the Grid, as well as examples of exponents of any dimension where possible. The provision of such (agreed) examples was, however, well beyond the remit of this Project. Nevertheless, and despite this level of disagreement, it was clear that completion of the Grid by individuals or groups could facilitate useful comparisons of results and discussions of the reasons for the different perceptions. This in itself could lead to enhanced understanding of the CEF and the categories in the Grid.

Correlations among analysts on item CEF levels were significant, ranging from a moderate .49 to a high of .78. Individual analysts' agreement with actual item and task levels were only moderate, in the .50s and .60s. Such relatively modest correlations show again the need for training, team discussions and team decisions when inputting data to the Grid.

Analyses were conducted separately for dimensions pertaining to items and those pertaining to texts. Although Chi-squares were calculated to test the strength of associations between dimensions and CEF levels, most did not meet the necessary levels of expected cell frequencies and so results can only be seen as tentative. Nevertheless, the Project team considered that they enable the development of initial hypotheses about relationships, which would have to be falsified in further research using a greater number of items and tasks.

Item type showed no association with CEF levels, and the operations judged to be tested by items varied considerably across analysts, sometimes reaching significance, sometimes not. Analysts 3 and 5 agreed that the commonest operation was "Recognise and retrieve explicit detail", whereas Analyst 2 found "Recognise and retrieve explicit main idea/ gist" to be commonest. Analyst 1 disagreed but agreed with Analyst 4 that the most frequently occurring operation was "Evaluate implicit text structure/ connections between text parts". Overall there was only moderate agreement among analysts as to what was being tested by individual items. There were also substantial differences among analysts as to which operations they identified in the various items. This is in line with findings in the literature on reading in a foreign language (see Alderson 2000) but it presents considerable difficulties for those who wish to claim that CEF levels can be distinguished by operations or "skills", and it underlines the finding that it is in fact rather difficult to reach agreement on what operations are required by any given item, at any CEF level.

However, when the CEF levels were grouped into three (A, B and C) the results showed a tendency for items at lower levels to be more focused on retrieving explicit information from texts, while at higher levels inferring from and evaluating texts became more prominent, and items tended to deal more with implicit information. Thus, some hope is provided by rather coarser-grained analyses at three CEF levels, reinforcing the desirability of further research using larger samples of texts and items in order to explore possible relations, especially if the analysts are trained in advance and discuss their analyses among themselves before reaching final decisions.

With respect to text characteristics, no significant association was found between the CEF level of a text/ task and authenticity, domain, grammar, text source, broad discourse type or narrow discourse type, topic or degree of abstractness of content. The only dimension that showed a significant association was vocabulary. Interestingly greater agreement among analysts was reached in Phase Three (69%) than had been reached in Phase Two using Grid 2 (where vocabulary was an open-ended category and agreement only reached 51%). Unlike in Phase Two, however, the number of words in a text or a task showed no clear association with CEF levels.

In conclusion, although Grid 3 was applied to 77 items belonging to 16 tasks, it must be stressed that more extensive research using the Grid is needed before solid conclusions can be reached about the relationship, or lack of it, between the dimensions of the Grid and CEF levels. Our results can only be considered to be suggestive, given the limited time available for this Project.

Nevertheless, the Grid is a useful instrument for the description of test items and tasks in terms of the CEF. Inter-analyst agreement is at times quite promising, but can clearly be improved by training. The relation between the dimensions in the Grid, and individual CEF levels is, however, not yet very obvious. Relatively few dimensions showed any significant association. However, based on the limited range of tasks and items analysed in this Project, one could hypothesise that variables like vocabulary, number of words in texts and tasks, operations and domain may well bear some relationship to CEF levels, whereas text source, discourse type and authenticity are less likely to have a clear relation with CEF levels. However, the collection of much more extensive data using Grid 4 is recommended as a priority for future research, to confirm or disconfirm these hypotheses.

The analysis of test specifications

In addition to analysing test tasks and items, colleagues were asked to analyse the specifications of tests to which they had access, using the dimensions of Grid 3, without the need to input data into the database. The aim of this procedure was to see to what extent tests produced by different examining bodies at the same level of the CEF agreed in content and specifications.

The tests whose specifications and related documents were examined were:

Cambridge ESOL: PET, FCE, CPE (including confidential documents), KET, CAE (publicly available documents only);
The Catalan Official Schools of Languages Exams Levels B1 and B2 (Elemental and Aptitud) - all languages;
Profile Tests Dutch as a Second Language;
Finnish National Certificates: all languages;
Certificats de français;
Zertifikat Deutsch.

A detailed analysis of the results is to be found in Appendix 12 and Appendix 13 contains a compilation of test specifications, by CEF Level.

The first conclusion is that the Grid turned out to be a useful instrument to describe and analyse the test specifications examined. It could therefore perhaps be used to analyse the diverse practices in language testing across Europe and it offers a tool for describing relations between specifications and the CEF. Secondly, we found – beside general similarities - many differences in the way test specifications dealt with descriptions of the characteristics of input texts and items for listening and reading, and in the terminology used. Thirdly, however, and importantly, there appear to be no systematic differences in the test specifications examined, in terms of most of the dimensions included in Grid 3, as CEF level changes. The specifications examined barely distinguish among CEF levels in terms of content.

The specifications analysed do not seem to be based on a theoretical construct, on how the language to be tested is understood. It appears that some specifications have been written focusing on the details of exam format and length for a particular level, without seriously considering language proficiency as a whole. This may be the reason why there is a lack of systematic and clear use of terminology, and also why there is a lack of uniformity of style and approach across levels.

Most importantly for this Project, there is very little information on how different dimensions may affect difficulty, or how the dimensions may vary across CEF levels. A common understanding of the specifications by item writers seems to rely in most cases on exemplification (previous exams) and expertise. This suggests the need – in addition to item writer training – to provide illustrative examples for the Grids produced in order to guarantee a common understanding of whatever terms or labels are used.

Recommendations: Proposals for further work on the Grid

The Project team feels that the Grid is a very useful tool, which is a valuable outcome of the Project, and it should be made as widely available as possible. It is felt that it would be of value specifically for the EU-funded Item Bank Project, as it would enable those who are analysing test items to categorise them appropriately, prior to their use in an item bank linked to the CEF. Members of the Project team would also like to use the Grid in their own various projects, in Finland, France, Germany, the Netherlands, Spain, and UK.

The Project team recommends the following further activities:

1. More work needs to be done on Listening as it was not covered as thoroughly as Reading in the Project.

2. A User's Guide needs to be developed, with rationale, advice on the use of the Grid (e.g. in teams rather than individuals), and misuse (e.g. belief in Estimated Level as being Real Level) together with an emphasis on the need for empirical trialling and calibration. It should be pointed out that the elements in the Grid do not represent a full test specification, but are an essential part of them. Other elements might be listed which are needed in a test specification, to be used in conjunction with the Manual and the CEF.

3. A training component needs to be added to the Grid, with sample pre-analysed items, by Dimension ("View an example"), and some explanation; a Glossary of Terms, hyper-linked from the Grid, would also be useful. In the Training mode, the actual item/ task level could pop up as feedback after all fields have been completed. A further addition could be to allow previous analyses (by experts) to be shown in comparison with the trainee's analysis, dimension by dimension.

4. An annotated bibliography could be provided, to encourage further use and development of the Grid in specific contexts for specific purposes (hopefully with information on such developments passed on to the Project team).

5. A "Comments on Item Quality" text entry box might be added, at the end of each Item input section (optional, not obligatory).

6. Hyperlinks from the Grid could be made to specific sections of the CEF, to the Manual, the Reference Supplement, Threshold, Profile Deutsch, Niveau Seuil, etc.

7. Data analysis tools should be developed to facilitate the statistical analysis of the results of using the Grid.

8. Clear guidance is needed concerning the language of texts and tasks at each level, in terms of grammar and vocabulary but maybe also sociolinguistic and pragmatic aspects of the development of foreign language proficiency.

9. Text and item interaction should continue to be explored by a) identifying which text section(s) the questions refer to, b) where in the text is the source for various multiple-choice options, c) how directly can the information for a) and b) be derived from the text, and so on. Initial agreement on this interaction and how it develops during discussion should be explored and the judgements should be compared to the results of item analysis.

10. A comparison is needed between what test specifications claim is being tested with an analysis of a large sample of actual test tasks and items.

11. A companion Guide for Item writers on how to link test items to the CEF would be of benefit.

12. Guidance on the creation of item banks based upon the results of the Grid might also be a useful facility.

Finally, it would be very useful to develop a mechanism to promote the co-ordination of such work throughout Europe.

Conclusions

The Dutch CEF Construct Project has developed a framework, based on the Common European Framework (CEF), for analysing language test items, test tasks and test specifications, in order to help test developers relate their examinations to the CEF. This framework has been developed as a web-based form (which we call a Grid - www.ling.lancs.ac.uk/cefgrid), which is completed by analysts, and whose data goes into a database that facilitates the analysis of results, from the point of view, *inter alia*, of the amount of agreement among analysts on the content of the test items, tasks etc.

It may not sometimes be appreciated clearly enough what a big difference it has been to move from the rather general description of language skills in syllabuses and test specifications to standards-based assessment. This is based on the specification of performance levels, which provide qualitative descriptions of the intended distinctions between adjacent levels of performance. Rather than rating test takers as “advanced”, “intermediate” or “beginner”, or as “good”, “satisfactory” or “fail” or in some other similar way, test takers receive a level rating, which describes what a person at that particular level can do with the language. This has been a challenge to psychometrics and test development and rating alike.

The Common European Framework (CEF) has received a lot of attention in the European language teaching community. One indication is that the CEF has been translated into some twenty languages. The CEF is increasingly referred to across Europe, and there is an urgent need to illustrate the levels of the CEF with calibrated test items. European Commission funding has recently been approved for a project that will attempt to construct an item bank that could serve as a common operational tool to enable the linking of national tests and examinations to the CEF.

In the domain of assessment, there have been calls on several occasions for the Council of Europe to take a more active role in assisting examination providers in their efforts to situate their

examinations within the Common European Framework, and in validating – in one form or another – language examinations that claim such linkage. It is in the interest of language examination providers to overcome difficulties in establishing valid and reliable links between the results of their systems and the levels of the CEF in order to make these links transparent to users of their language examinations.

A preliminary pilot Manual was produced by the Council of Europe in response to the need for guidance to assist examination providers to relate their examinations to the CEF. The Manual notes that the setting of standards related to reading and listening comprehension items is a particularly difficult challenge. In a sense, the Dutch CEF Construct project has worked on this problem and our work may be of some use, both in the Item Bank Project referred to above, and in the revision of the Manual. On the other hand, the Project's work benefitted from the work done on the Manual. The Manual and the accompanying Reference Supplement are intended to encourage a thoughtful approach to relating tests and examinations to the CEF. A set of procedures have been proposed in the Manual which help to increase awareness of the test or examination itself, to improve familiarity with the CEF, to standardise judgments and to analyse the validity of the claimed linkage. Specific caveats are issued at several points to avoid oversimplification and thoughtless application of rules of thumb.

This advice in the Manual applies with the same force to the kind of work that this Project has done in relating listening and reading items to the CEF. Our work has shown how complex this is. We believe that we have made some definite progress but also recognise that much more could and should be done. Like the authors of the Manual, we urge the users of the grids to use them thoughtfully. There is a need to apply the grids to a number of different kinds of texts and items, to review with colleagues the classifications, to discuss the possible reasons for differences, to compare estimated CEF-level ratings to empirical difficulty indices, etc. In addition, if we wish our ratings of listening and reading items to have international validity, it is necessary to have items rated by an international panel.

It was not the object of this Project to conduct extensive research into what makes reading and listening tests difficult, but rather to seek to develop, on the basis of the CEF but by complementing it where it appeared necessary, an instrument based on a theoretical framework which would enable test developers and item writers to produce test items that corresponded to the constructs elucidated in the CEF, and which could be calibrated to the different CEF levels. The empirical research that we have been able to conduct in the limited time available suggests that the CEF does not provide sufficient guidance to enable item writers to develop tests at specific levels of the CEF. However, it is clear that this tentative conclusion needs to be replicated in much larger scale studies, which can probably only be undertaken once there exists a body of tests and tasks which have been developed explicitly to correspond to the CEF, and which have been empirically linked to the CEF. At present, relatively few such tests exist. Nevertheless, it is important to undertake more extensive research into the usefulness of the Grid for the characterisation of texts and tasks at the different levels of the CEF.

Indications from our necessarily limited research are that the dimensions of the Grid (and thus of the CEF) do not distinguish among the levels of the CEF. Indeed, we proposed a procedure in Phase One, whereby content analysis of test texts and items should proceed hand in hand with

empirical investigations of difficulty, and empirical standard-setting procedures, as outlined in the Manual. We repeat those recommended procedures here:

- Describe the text and items using the dimensions of a classification system (The Frames and Grids).
- Make a guess at the level of an item (guided by the classification system and the CEF scales), leading to an estimated CEF-level.
- Pretest the items thus labelled, describing in detail the characteristics of the pilot sample.
- Calibrate the items.
- Do standard setting to set the boundaries of the levels on the scale coming from the calibration.
- Assign a psychometric level to the items.
- Assign a definitive level to the items. An item can only be assigned to a definitive level if the psychometric level falls within the band of the estimated level (in other words if the estimation based on the **analysed content** is comparable with the **psychometric** value).

In short, the identification of separate levels in the CEF is at least as much an empirical matter as it is a question of the content of the tests as determined by test specifications or as identified by our Grids. However, it is the suspicion of the Project team that it will be necessary to examine the linguistic characteristics of texts and items in much more detail than has been possible in this Project, if adequate characterisations of the content and construct of tests of language proficiency at the different levels of the CEF are to be determined. This is likely to involve

a) identifying tests and tasks which have been incontrovertibly scaled on the CEF,

b) developing measures of the linguistic features of texts and tasks which previous research has shown to be relevant to defining difficulty, independently of the CEF (see, for example, Buck et al, 1997, and Shiotsu and Weir, 2004) and

c) applying such measures experimentally to the texts and tasks identified in a) to see to what extent analysis of the linguistic features of such texts and tasks can predict CEF levels. This is likely to be a rather extensive project, and we therefore recommend that it be conducted first for reading, for two languages only - English and French - and that, if successful, the research be later extended to listening for the same languages, and only then in a third phase to other languages.

However, lest this conclusion seem unduly pessimistic, we wish to affirm that the Project has developed an instrument, whose latest version, Grid 4, provides a promising framework to enable the characterisation of test items and tasks. It has been positively evaluated by a number of groups as this report was being written, in particular members of the test development team of the Budapest Business School, participants on the Language Testing at Lancaster 2004 summer school, and analysts associated with related projects (see the report of the use of the Grid in the DESI project, Appendix 14).

References

Alderson, J. C. (2000) *Assessing Reading*. Cambridge Language Assessment Series. Cambridge: Cambridge University Press

Alderson, J. C., Clapham, C. and Wall, D. (1995) *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. and Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Buck, G., Tatsuoka, K. and Kostin, I. (1997) The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, Vol. 47 (3). pp 423-466.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2003) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Preliminary Pilot Version of The Manual.* Strasbourg: Language Policy Division, The Council of Europe.

Davidson, F. and Lynch, B. (2002) *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.

Shiotsu, T. and Weir, C. J. (2004) The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. Unpublished manuscript.

van Ek, J. A. and Trim, J. L. M. (1998) *Threshold 1990*. Cambridge: Cambridge University Press.

van Ek, J. A. and Trim, J. L. M. (1998) *Waystage 1990*. Cambridge: Cambridge University Press.