UCLES-RCEAL Funded Research Projects

The following projects are described here:

Project (1) continues current activities on the Cambridge Learner Corpus (in collaboration with Prof. Ted Briscoe and ILexIR). It comprises three subprojects: (1.1) continues the current hypothesis testing and the search for criterial features at each proficiency level and for L1 transfer effects (John A. Hawkins, PI); (1.2) and (1.3) expand on the results found so far with more detailed follow-up projects (Dora Alexopoulou, PI for (1.2) and Teresa Parodi, PI for (1.3)).

Project (2) (Henriëtte Hendriks, PI) makes use of CLC corpus data but goes beyond it with a more functionally-oriented study involving discourse structure and style. This project is of special significance for the higher proficiency levels, B2 - C2.

Project (3) (with several PIs, cf. below) focusses on the visibility and dissemination of EPP-related research through major conferences, workshops and the EPP website.

Project (4) involves the search for much-needed additional corpus data, without which many current hypotheses cannot be properly tested (Dora Alexopoulou, PI).

PROJECT 1

The Cambridge Learner Corpus (CLC) Project
Professor John Hawkins, overall PI

(a) Project Summaries, Research Goals and Outcomes
Project (1) continues current activities on the Cambridge Learner Corpus (in collaboration with Prof. Ted Briscoe and ILexIR). It comprises three subprojects: (1.1) continues the current hypothesis testing and the search for criterial features at each proficiency level and for L1 transfer effects (PI, John A. Hawkins); subprojects (1.2) and (1.3) expand on the results discovered so far with more detailed follow-up projects (Dora Alexopoulou, PI for (1.2) and Teresa Parodi, PI for (1.3)).

Subproject (1.1): Hypothesis Formulation and Testing (PI, John Hawkins)

Project Summary
Professor Hawkins has identified a set of 20 lexical and grammatical areas that are promising initial candidates for criterial feature identification and transfer effects at different proficiency levels (cf. §1.1). These properties are searchable, given the CLC in its current tagged, parsed and error-coded form. The hypotheses are informed by acquisition studies hitherto, by observed transfer effects, and by psycholinguistic metrics of complexity. Some of these hypotheses have been tested already, and results to date are summarized in §1.2. The data resulting from this hypothesis-testing have led to further hypotheses that must be tested. The next phase of this project has three goals (cf. §1.3): to identify further hypotheses for testing and to actually test them (§1.3.1); to incorporate these and prior results into a revised set of general findings and predictions for the EPP (§1.3.2); and to pursue two detailed subprojects and incorporate their results into the general findings and predictions, involving on the one hand nouns and determiners, i.e. noun phrases (Subproject (1.2), PI Dora Alexopoulou), and on the other hand verbs and clausal syntax (Subproject (1.3), PI Teresa Parodi).

Research Goals and Outcomes
1.1 <u>20 lexical and grammatical areas for hypothesis testing</u>
The 20 lexical and grammatical areas are divided into Lexicon, Morpho-Syntax, Syntax, and Overall Metrics. The error codes RV, RN, etc, are taken from the CLC coding system, and example sentences given here are taken, wherever possible, from the Cambridge University Press working document entitled "The <#S>Compleat ǀComplete</#S> Learner Corpus", dated 10/11/05. The proficiency levels are as follows:

A1 Breakthrough (no data in the CLC)
A2 Waystage
B1 Threshold
B2 Vantage
C1 Effective Operational Proficiency
C2 Mastery.

LEXICON

1.1.1 Lexical Choice Errors: Noun (N) and Verb (V)

RN	Replace noun	Have a good travel!
FFN	False friend noun	It was an interesting history
RV	Replace verb	I existed last weekend in London
FFV	False friend verb	I passed last weekend in London

Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors (quantify R and FF errors separately and together);

(II) items subject to error will correlate inversely with native speaker frequencies in the British National Corpus (BNC), i.e. the more errors the lower the frequency of the N or V in the BNC;

(III) error rates will vary with L1: e.g. genetically distant L1s (Chinese, Japanese, Korean) will exhibit more lexical choice errors than for L1s that are genetically close to English or lexifier languages or closely related to lexifier languages (German, French, Spanish respectively).

1.1.2 Lexical Occurrences: Noun (N) and Verb (V)

Develop a "lexical frequency profile" for A2-C2, quantifying:

(i) the number of Ns and Vs used more than X times at each of A2-C2;

(ii) the increases in N and V usage through A2-C2;

(iii) the correlations between N and V usage in (i) and (ii) and corresponding frequencies in the BNC.

Hypotheses:

(I) there will be a positive correlation between N/V usage and increases in usage through A2-C2 and BNC frequencies, i.e. (i) and (ii) should correlate with BNC frequencies, the more uses the higher the BNC frequency;

(II) it should be possible to define a "lexical frequency threshold" for each level: B1 admits items (N/V) of frequency X in BNC, but not X-1, B2 admits X-1 but not X-2, etc.

(III) lexical occurrence rates will vary with L1: e.g. genetically distant L1s (Chinese, Japanese, Korean) will exhibit fewer productive lexical occurrences than for L1s that are genetically close or lexifier languages or closely related to lexifier languages (German, French, Spanish respectively), lesser productivity being measured (a) in quantities of Ns and Vs at each proficiency level and (b) in higher frequencies of BNC occurrence required for each level.

1.1.3 Lexical Choice Errors: Adjective (J) and Adverb (Y)

RJ	Replace adjective	The afternoon was very bored
FFJ	Adjective false friend	My ancient girlfriend
RY	Replace adverb	He stared at her intensively
FFY	Adverb false friend	

Hypotheses:

Same as (I)-(III) in §1.1.1.

1.1.4 Lexical Occurrences: Adjective (J) and Adverb (Y)
Develop the same lexical frequency profile as in §1.1.2.
Same hypotheses (I)-(III) as in §1.1.2.

1.1.5 Verb Co-occurrence Errors
Examine verb-preposition co-occurrences.

RT	Replace preposition	When I arrived at London
MT	Missing preposition	I gave it John
UT	Unnecessary preposition	I told to John that ...

Caution: distinguish lexically defined V-Prep sequences from free-standing (and semantically independent) prepositions.
Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors;
(II) co-occurrences subject to error at each of A2-C2 will correlate inversely with native speaker frequencies in BNC, i.e. the more errors the lower the frequency of the V-Prep co-occurrence in BNC;
(III) error rates will vary with L1, e.g. genetically distant L1s (Chinese, Japanese, Korean) will exhibit more verb co-occurrence errors than for L1s that are genetically close or lexifier languages or closely related to lexifier languages (German, French, Spanish respectively); in addition V-Prep co-occurrence errors translated from particular L1s can be identified ("wait on" rather than "wait for" among German learners of English) and hypothesis (I) can be applied to this subset of V-Preps.

1.1.6 Verb Co-occurrence Uses
Develop a lexical frequency profile for V-Prep co-occurrences.
Hypotheses:

(I) there will be a positive correlation between increases in V-Prep co-occurrence usages through A2-C2 and BNC frequencies, i.e. the more V-Preps the higher their BNC frequency;
(II) it should be possible to define a lexical co-occurrence threshold for each level: B1 admits V-Prep co-occurrences of frequency X in BNC, but not X-1, B2 admits X-1 but not X-2, etc.
(III) V-Prep co-occurrence uses will vary with L1, e.g. genetically distant L1s will exhibit fewer productive V-Prep co-occurrences than for L1s that are genetically close, etc, cf. §1.1.5 (III).

MORPHO-SYNTAX

1.1.7 Infinitival Complement of Verbs: Errors

FV Wrong verb form (NB! immediately following a higher V or Aux)
I suggest to look for another restaurant (instead of: looking)
We must bearing in mind that ... (instead of: bear)

Hypothesis:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors.

1.1.8 Number Agreement NP-internally and on Verbs: Errors

AGN	Noun agreement error	One of my friend
AGD	Determiner agreement error	I enjoy these job
AGQ	Quantifier agreement error	In another circumstances
AGV	Verb agreement error	The three birds is singing

Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors;

(II) verb agreement errors ≥ NP-internal errors at each level;

(III) verb agreement errors will be greater for non-adjacent (head of) subject NP and verb than for adjacent, e.g.

The three birds I was telling you about is singing ≥ The three birds is singing

The three birds yesterday was singing ≥ The three birds was singing

SYNTAX

1.1.9 Missing Determiner Errors

MD (a)	Missing determiner	I have car
MD (the)		I spoke to President

Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors;

(II) error rates will be greater (or equal) for L1s without articles (Japanese, Korean, Chinese, Russian, Polish, Turkish) than for L1s with (Spanish, French, German).

1.1.10 Determiner Choice Errors

RD (e.g. the → a)	Replace determiner	Have the nice day

Same hypotheses (I) and (II) as in §1.1.9.

1.1.11 Word Order Errors: Verb-Object Separation

W	I have also two cats
	She liked a lot her aunt

Identify V-XP-NP, where NP = direct object of V and XP = AdvP or PP or NP other than indirect object of V (or direct object of V).

Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors;

(II) error rates will be greater (or equal) for L1s with VO word order and V-fronting (/movement/raising), e.g. Spanish, French, than for L1s with OV and no parallel rule (Japanese, Korean).

1.1.12 Word Order Errors: Genitive Positioning

W	The room of my son (for: my son's room)

PossP-N alternates with N-PP (P=of) in English, with subtle conventions distinguishing them. Identify errors involving each.

Hypotheses:

(I) error rates will decline from A2-C2, i.e. the higher the proficiency level, the fewer (or equal) errors;

(II) error rates will be greater in the direction of PossP-N for L1s with Genitive-Noun orders (Japanese, Korean, Chinese, Turkish) and in the direction of N-PP for L1s with Noun-Genitive orders (French, Spanish, Russian, Polish).

1.1.13 Relative clause Uses and the Keenan-Comrie Noun Phrase Accessibility Hierarchy

Identify relative clauses whose heads function as Subject (SU), Direct Object (DO), Indirect Object/Oblique (IO/OBL) or Genitive (Gen) within the immediately adjacent (highest) S of the relative clause that is a sister to the head.

AH: SU > DO > IO/OBL > GEN

the professor [that wrote the letter]	SU
the professor [that the student knows]	DO
the professor [that the student showed the book to]	IO/OBL
the professor [whose son the student knows]	GEN(-DO)

Hypotheses:

(I) at each A2-C2 level the relative frequencies of relative clauses will follow AH (SU ≥ DO ≥ IO/OBL ≥ GEN);

(II) the distribution of usage frequencies will shift from higher to lower positions of the AH from A2-C2, gradually approximating the relative frequencies for native speakers in the BNC.

1.1.14 Relative Clause Errors and the AH

Identify those relative clauses that contain errors in the form of a resumptive pronoun in SU, DO or IO/OBL position.

the professor [that he wrote the letter]	SU
the professor [that the student knows him]	DO
the professor [that the student showed the book to him]	IO/OBL

Hypotheses:

(I) retained pronouns, if they occur, will be favored in IO/OBL ≥ DO ≥ SU positions, i.e. reverse AH effect;

(II) retained pronouns will be especially favored in those L1-L2 pairs when the L1 uses them gramatically (e.g. Persian, Mandarin).

1.1.15 Relative Clause Uses and the Link to the Subcategorizer (Verb/be+Adjective)

Identify the distance from the head of the relative to its subcategorizor (verb), for those relative clauses in which the head is an argument of a verb and in a GR with it. The head should be a SU, DO or IO only and should not be properly contained within any higher NP that contracts these relations with the subcategorizer. The head (H) and subcategorizer (S) are co-indexed in the following examples:

(i) the professor*i* [that the student knows*i*]

(ii) the professor*i* [that I believe is clever*i*]

(iii) the professor*i* [that I believe that the student knows*i*]

The distance between co-indexed items is the Head-Subcategorizer Domain (HSD), measured in words (terminal elements):

(i) HSD = 5 (include head and subcategorizer)

(ii) HSD = 6

(iii) HSD = 8

Hypotheses:

(I) HSD averages will increase from A2-C2, i.e. they will be greater (or equal) at each higher proficiency level.

(II) at each A2-C2 level relative frequencies of usage will correlate inversely with HSD sizes, i.e. the larger the HSD, the less (or equally) frequent;

1.1.16 Wh-movement Uses and the Link to the Subcategorizer (Verb/be+Adjective)

Select fronted main clause Wh-words that are arguments of subcategorizers, i.e. in a (SU, DO or IO) GR with some verb. Relevant Wh-phrases are: who, whom, what, which, which N, what N, whose N. The Wh-word (W) and subcategorizer (S) are co-indexed in the following examples:

(i) who*i* does the student know*i* ?

(ii) who*i* do you believe is clever*i* ?

(iii) who*i* do you believe that the student knows*i* ?

The distance between co-indexed items is the Wh-Subcategorizer Domain (WSD), measured in words (terminal elements):

(i) WSD = 5 (include Wh and subcategorizer)

(ii) WSD = 6

(iii) WSD = 8

Hypotheses: same as (I) and (II) in §1.1.15, with WSDs replacing HSDs.

1.1.17 Tough Movement Uses and the Link to the Subcategorizer (Verb)

Select structures in which a (subject) NP is separated from its subcategorizer (in a non-subject GR) by a "Tough" predicate. The Tough Subject (T) and subcategorizer (S) are co-indexed in the following examples:

(i) this book*i* is easy to read*i*

(ii) this book*i* is easy for me to read*i*

(ii) this book*i* is easy for me to persuade Harry to read*i*

The distance from T to the subcategorizer is a TSD, measured in words (terminal elements):

(i) TSD = 5 (include T and subcategorizer)

(ii) TSD = 7

(ii) TSD = 10

Hypotheses:

(I) the number of Tough structures will increase from A2-C2, i.e. more (or equal) Tough structures, the higher the proficiency level;

(II) TSD averages will increase from A2-C2, i.e. they will be greater (or equal) at each higher proficiency level.

(III) at each A2-C2 level relative frequencies of usage will correlate inversely with relative TSD sizes, i.e. the larger the TSD, the less (or equally) frequent.

1.1.18 Raising Structure Uses and the Link to the Subcategorizer (Verb/be+Adjective)

Select structures in which a (subject) NP is separated from its subcategorizer (in a subject GR) by a "Raising" predicate (cf. Postal 1974 for a full listing). The Raising Subject (R) and subcategorizer (S) are co-indexed in the following examples:

(i) John*i* seems to enjoy*i* tennis

(ii) John*i* is likely to win*i*

(iii) The job*i* appears to be easy*i*

Hypotheses:

(I) the number of Raising structures will increase from A2-C2, i.e. more (or equal Raising structures, the higher the proficiency level;

(II) the relative frequencies of the different Raising predicates at each A2-C2 level will correlate positively with relative frequencies in the BNC.

OVERALL METRICS

1.1.19 Overall Error Counts

Quantify ALL errors per A2-C2 level for each L1-L2 pair.

Hypotheses:

(I) overall error rates will decline from A2-C2, i.e. fewer (or equal) errors at each higher proficiency level;

(II) overall error rates will correlate positively with the degree of typological and genetic distance between L1 and L2, i.e. the greater the distance, the more (or equal) errors in L2.

An Initial Scale of Typological and Genetic Distance from English (using basic word order and head ordering as (the only) typological features and Indo-European (IE) vs non-IE, and divisions with IE, as (the only) genetic features)

1	2	3	4	5	6
English	**Spanish**	**Russian**	**Vietnamese**	**Chinese**	**Japanese**
SVO	**French**	**Polish**	SVO	SVO/SOV	**Korean**
Germanic with	SVO	SVO	non-IE	non-IE	**Turkish**
signficiant	Romance	Slavic			SOV
Romance	IE	IE			non-IE
lexicon					
IE	**German**				
	SVO/SOV				
	Germanic				
	IE				

1.1.20 Overall Syntactic Complexity Metric

Quantify mean sentence and clause (S) complexity scores per A2-C2 level for each L1-L2 pair, using words (terminal elements).

Sentence: quantify word totals from full stop to full stop

Clause: quantify the number of words dominated by each S node (excluding words that have an additional intervening S node)

Hypotheses:

(I) sentence and clause complexity will increase from A2-C2, i.e. the mean complexity will be greater (or equal) for each higher proficiency level;

(II) overall complexity scores will correlate inversely with the degree of typological and genetic distance between L1 and L2, i.e. the greater the distance, the less (or equal) the overall complexity in L2.

1.2 Summary of Results from Hypothesis Testing to Date

1.2.1 Determiner errors

In order to investigate hypothesis (9) (§1.1.9) preliminary work has quantified error rates for missing determiners (specifically, zero in place of "the" and "a"). It was expected that error rates would decline from A2-C2 and that error rates would be greater (or equal) for L1s without articles. Results have shown that the general trends support the hypotheses. However further investigation is required since the existence or absence of articles in L1 can only partially predict L2 production. For instance, speakers with L1s that have articles are sometimes producing article omission mistakes. Also, there are differences in the error rates for L2 learners whose L1s lack a definite article. The results are shown in Tables 1.2.1.a and 1.2.1.b. These findings have motivated subproject 1.2.

Table 1.2.1.a Missing Determiner Error Rates for L1s with Articles

	Missing "the"				
	A2	***B1***	***B2***	***C1***	***C2***
French	4.76	4.67	5.01	3.11	2.13
German	0.00	2.56	4.11	3.11	1.60
Spanish	3.37	3.62	4.76	3.22	2.21

	Missing "a"				
	A2	***B1***	***B2***	***C1***	***C2***
French	6.60	4.79	6.56	4.76	3.41
German	0.89	2.90	3.83	3.62	2.02
Spanish	4.52	4.28	7.91	5.16	3.58

Table 1.2.1.a shows missing determiner error rates for "the" and "a" at all proficiency levels for French, German and Spanish as first languages. All three languages have an article system. The figures indicate the percentage of errors with respect to the total number of correct uses. For instance a percentage of 10.0 would indicate that a determiner was omitted 1 in every 10 times that it should have appeared. We see generally low error rates for these languages, without significant deviation between levels.

Table 1.2.1.b Missing Determiner Error Rates for L1s without Articles

	Missing "the"				
	A2	***B1***	***B2***	***C1***	***C2***
Turkish	22.06	20.75	21.32	14.44	7.56
Japanese	27.66	25.91	18.72	13.80	9.32
Korean	22.58	23.83	18.13	17.48	10.38
Russian	14.63	22.73	18.45	14.62	9.57
Chinese	12.41	9.15	9.62	12.91	4.78

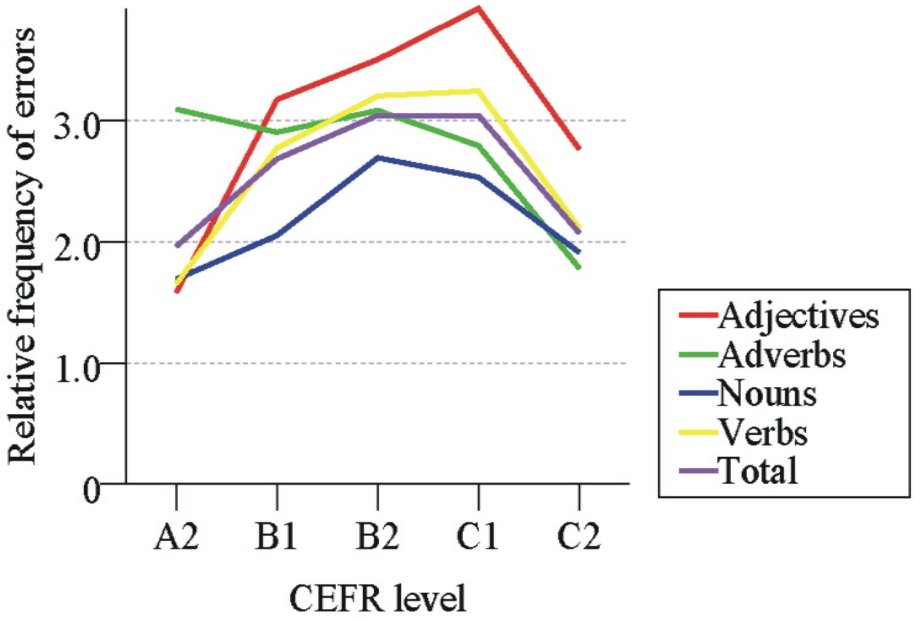
	Missing "a"				
	A2	***B1***	***B2***	***C1***	***C2***
Turkish	24.29	27.63	32.48	23.89	11.86
Japanese	27.66	25.91	18.72	13.80	9.32
Korean	22.58	23.83	18.13	17.48	10.38
Russian	14.63	22.73	18.45	14.62	9.57
Chinese	4.09	9.20	20.69	26.78	9.79

Tables 1.2.1.b shows missing determiner error rates for "the" and "a" at all levels for Turkish, Japanese, Korean, Russian and Chinese as first languages. These languages do not have an article system. There is a general linear improvement, i.e. a decline, in error rate across levels (from left to right). However, Chinese shows an inverted U-shaped progression (especially in the case of missing 'a'). The cause for this needs further investigation.

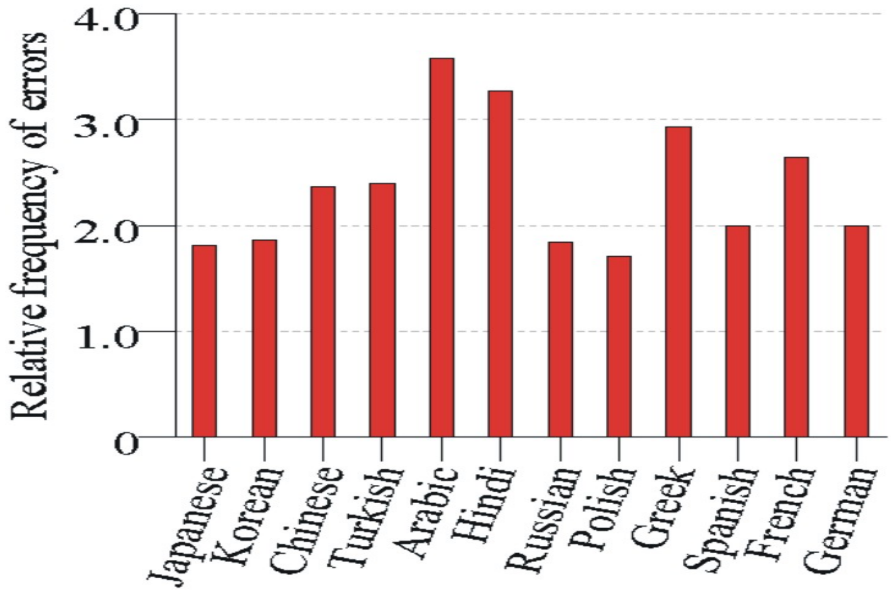
1.2.2 Lexical choice errors

Lexical choice errors (sections i and iii of hypotheses (1) (§1.1.1) and (3) (§1.1.3)) were investigated to see whether error rates in the use of the correct choice of noun, verb, adjective, and adverb (lexical choice errors) would decline from levels A2 to C2, and whether error rates varied with L1, with genetically distant L1s showing higher error rates than genetically related or lexifier languages. The research found that error rates rose from A2 to B2/C1 before falling again to C2. The C2 error rate was higher than A2 error rate but lower than the C1 error rate. There was no obvious relationship between error rate and type of L1.

Graph 1.2.2.a Relative Frequency of Lexical Choice Error Rates for various parts of speech plotted against CEFR level for different parts of speech (all languages combined).



Graph 1.2.2.b -- Average Relative Frequency of Lexical Choice Error rates across Levels for Each Language



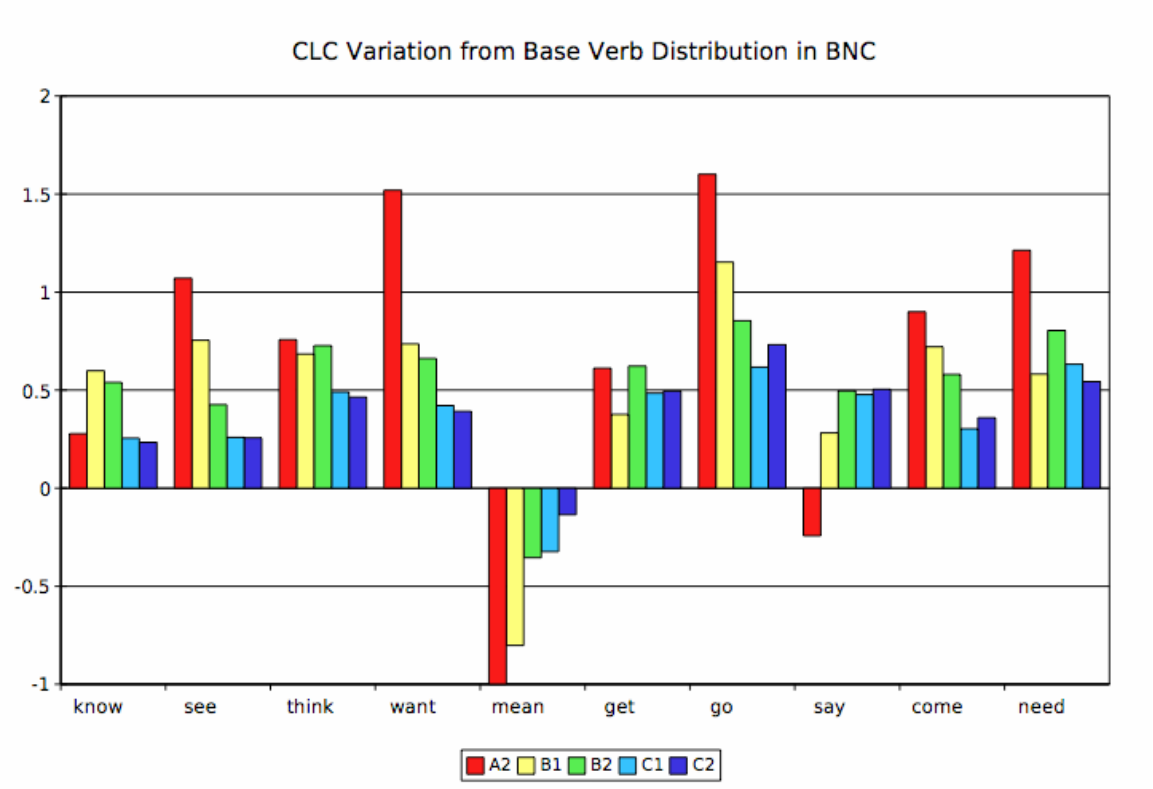
1.2.3 Lexical occurrences

Lexical occurrences within the CLC have been investigated and where possible analyzed with respect to a standard British English corpus---the British National Corpus or BNC (see hypotheses (2) (§1.1.2) and (4) (§1.1.4)). A detailed analysis was performed on the base forms (present tense, un-inflected) of lexical verbs within the CLC, with the following results:

(i) *Distribution of base verbs within the CLC versus the BNC*. Results showed a general trend which indicated that higher levels of proficiency (A2 to C2) increasingly approximate to the BNC distribution. The analysis revealed that common verbs are overrepresented in the CLC and it also highlighted those verbs which are underrepresented. This line of research, although promising, would benefit from data skew control. It is important to control for the use of verb forms in the questions that led to the students' answers.

ii) *Ratio of lexical errors to lexical instances of base forms of verbs*. Results across levels generally showed an inverted U-shaped learning curve that might be attributed to increasingly more complex verb usage within the mid-range levels.

Graph 1.2.3.a variation (expressed as a ratio of relative frequency) of the base form of lexical verbs in the CLC from the BNC. Bars above the zero line indicate an over-use in the CLC as compared to the BNC and bars under the zero indicate under-use.



1.2.4 Verb co-occurrences. A preliminary investigation into verb co-occurrence and complementation (hypotheses (5) (§1.1.5), (6) (§1.1.6) and (7) (§1.1.7)) has been carried out using a state of the art subcategorization frame extraction system. The system is capable of identifying and extracting 168 verb complementation patterns (frames) using a rule based classifier over the grammatical relation parser output of RASP. The analysis can quantify the similarity between verb usage in the CLC versus a standard British English Corpus. Table 1.2.4.a shows the similarity between the usage of the verb believe at different levels of proficiency against the usage in the BNC.

	B1	**B2**	**C2**
precision	100	100	85.7
recall	47.6	52.4	57.1
f-measure	64.5	68.8	68.6
intersection	0.65	0.69	0.73
rank correlation	0.74	0.75	0.81
KL distance	1.33	0.87	0.72

Table 1.2.4.a Quantifying usage similarity of the verb believe at B1, B2 and C2 proficiency as compared to the BNC. Precision indicates the proportion of all frames found in the CLC data that were also present in the gold-standard set of the frames from the BNC: in this case, at levels B1 and B2 all of the frames found were also present in the BNC, but at level C2 frames were encountered that we don't see in the BNC. Recall is the number of frames found in the CLC as a fraction of all the possible frames in the gold-standard: here we see learners making use of around half of the possible frames that are available for "believe". The f-measure is a weighted average of precision and recall. Intersection indicates the set intersection between the set of frames in the BNC and the set for that given level; rank correlation indicates how close the ranking of frames was within the two corpora (a rank correlation of 1 indicates identical rankings); KL distance measures the euclidean distance between the distribution of frames in the BNC and that of the given level (distributions are identical at a KL distance of 0).

1.3 Research Goals for 2007-08

The first research goal for 2007-08 is to expand the set of hypotheses to be tested beyond the list summarized in §§1.2.1-4. The additional areas to be tested are summarized in §1.3.1. The second goal is a more theoretical one. Some of our results have confirmed the general logic of the original hypotheses in §1.1. But some of the data go against some of the predictions. It is necessary to reconsider and in some cases reformulate some of the hypotheses for developmental stages and transfer effects, in the light of these data, cf. §1.3.2. The third goal involves going beyond the results of §1.2 with more detailed studies that can help us understand the errors and developmental stages better and that can contribute further to the theoretical work described in §1.3.2. This third goal is described in subproject (1.2) below, on specificity and determiners (PI, Dora Alexopoulou) and in subproject (1.3) on the morphosyntax of verbs (PI, Teresa Parodi).

1.3.1 Further hypotheses to be tested

The additional hypotheses to be tested will involve the following morpho-syntactic and syntactic areas:

- Number agreement NP-internally and on verbs, cf. §1.1.8
- Relative clauses, cf. §§1.1.13-15

- Wh-questions, cf. §1.1.16

Also to be tested are the hypotheses involving overall error counts:

- Overall error counts, cf. §1.1.19

1.3.2 Hypothesis formulation and reformulation

The hypotheses formulated in §1.1 are supported in many of the data sets summarized in §1.2. For example, the data involving determiner errors (cf. §1.2.1) provide a lot of support for predictions (I) and (II) in §§1.1.9-10. Where there are significant differences in incorrect determiner uses between levels (i.e. higher or lower ratios of incorrect uses as a percentage), the higher proficiency levels generally do have fewer errors, in accordance with hypothesis (I). And the first languages without articles (Japanese, Korean, Russian, Polish, etc) are in general associated with more errors than those with articles (Spanish, French, German), in accordance with hypothesis (II). The expanding verb co-occurrence frames of §1.2.4 also showed a linear progression from B1 to B2 to C2 (cf. the recall figures in Table 1.2.4.a).

The lexical choice errors of §1.2.2, however, show a different pattern: error rates first rise and then fall, resulting in an inverted U curve. Verb co-occurrence errors also increase from B2 to C2 in §1.2.4 (cf. the precision data in Table 1.2.4.a). This is a pattern that has been observed in first language acquisition in the learning of morphological irregularities (sing/sang versus walk/walked; goose/geese versus swan/swans; etc): children first use the correct forms, then get them wrong (as they learn the productive rules), and then get them right again.

The theoretical issue that this raises, for which current wisdom does not provide a general solution, is this: for which areas of second language acquisition should we expect a linear progression from A2-C2, along the lines of the determiner errors, and for which areas should we expect the inverted U curve? This issue will be researched in the light of results from the hypotheses tested in §1.2, incorporating the results from §1.3.1 and from subprojects (1.1) and (1.2). The hypotheses of §1.1 will be reformulated accordingly.

Subproject (1.2): Specificity and Determiners (PI, Dora Alexopoulou)

Project Summary

This project builds on the preliminary findings of the determiner error analysis (§1.2.1) and will examine the effect of the L1 on the acquisition of the English nominal system, with particular reference to definite and indefinite articles. The preliminary investigation confirmed the general hypothesis that learners with L1s lacking a definite article have more difficulty in acquiring the distribution of English articles and produce a higher rate of errors involving articles. Despite the correctness of this general hypothesis, the results indicate that the existence or absence of a definite article in L1 can only partially predict L2 production. First, speakers with L1s that have a definite article do produce article omission mistakes (though at a lower rate than speakers with L1s lacking a definite article). Second, there are significant differences in the error rates for L2 learners whose L1s lack a definite article. This project will investigate further the source and nature of these differences by focusing on the effect of finer interpretative distinctions in the use of the English determiner system. In particular, the project focuses on the interaction between specificity and the use of determiners in English.

Research Goals and Outcomes

2.1. Specificity and the English Determiner System

One way of illustrating specificity is by considering the two interpretations of *a house with a big balcony* in (1) (known in the literature as *epistemic* specificity).

(1) We are looking for *a house with a big balcony*.

a. Non-specific: We are looking (to buy) a house with a big balcony, but we have not identified one yet.

b. Specific: We are in the street trying to locate a house with a big balcony that we know exists.

As can be seen in (1) specificity is not unambiguously encoded by one form (e.g. the indefinite article) in English. Rather, both definite and indefinite articles can denote both specific and non-specific entities. This poses an important challenge for L2 learners, since the acquisition of the English determiner system is inseparable from the acquisition of specificity. So, the starting point for this project is the hypothesis in H1.

H1: Correct use/acquisition of the distribution of the definite and indefinite article relates to the acquisition of how specificity is encoded in English.

The project identifies four areas in which the acquisition of specificity will present L2 learners with specific challenges: (i) non-specific arguments of intensional predicates; (ii) predicative nouns; (iii) incorporated arguments; and (iv) misanalysis of the definite article as a specificity marker. More detailed hypotheses relating to these four areas are presented below.

2.1.1 Non-specific arguments of intensional predicates

Even languages with a full fledged determiner system allow determineless/bare noun phrases (e.g. "play football", "tigers are wild animals"). A common case crosslinguistically is that of non-specific indefinites as in: "I am looking for a flatmate" or "we want to buy a house" and a common mistake of L2 learners is the omission of the indefinite article in such examples. This tends to correlate with the omission of the indefinite pronoun one in exchanges like (2).

(2) a. Have you found flatmate?
b. Yes we found (one). He's moving in next Saturday.

The project will investigate such cases in relation to H1a:

H1a: Learners of English with L1s missing an indefinite pronoun like English *one* and/or missing a morphological distinction between the numeral (*one*) and the indefinite article (*a*) will systematically omit the indefinite article and the indefinite pronoun in contexts involving intensional predicates like *look-for* or *want*.

2.1.2. Predicative nouns

In many languages predicative uses of nominals necessarily involve a bare noun; L2 learners of English speaking such languages tend to omit the indefinite article in examples like (3) and (4):

(3) Jane is doctor.

(4) I had treated her as mother.

H1b: Learners of English with L1s using bare predicative nouns will produce higher rates of omission of the indefinite article in relevant contexts.

2.1.3. Incorporated arguments.

Crosslinguistically, incorporation of verb and noun is common (characteristically in languages lacking determiners). In languages with determiners, equivalent predicates are often expressed with bare nouns; eg. "play piano" (as opposed to the English "play the piano") or "have car" (as opposed to the English "have a car"---see section 1.1.9 for discussion of such errors in the CLC).

H1c: Speakers of L1s with systematic verb-noun incorporation will produce higher rates of article omission with the corresponding predicates in English.

2.1.4. Misanalyis of the defininite article as a specificity marker.

The cases discussed so far mainly involve errors relating to the incorrect use (omission) of the indefinite article. However, a common pattern, in particular for L2 learners lacking a definite determiner in their L1s, involves the misanalysis of the definite article as a marker of specificity rather than definiteness. The consequences of this are: (i) incorrect omission of the definite article in environments involving definite but non-specific phrases as in the case of generic nominals (5) and (6); and (ii) incorrect use of the definite article for specific indefinites as in (7) and (8). Example (7a) is pragmatically odd because the definite article "the" implies that there is only one student in our class, while world knowledge indicates that classes (normally) have more than one student. This uniqueness presupposition is absent in (7b) which involves the indefinite article. Similarly, L2 learners may produce (8a) with the intended meaning of (8b); both (8a) and (b) involve a specific nominal, but, "the wallet" in (8a) is also familiar or unique in the discourse while (8b) is novel and not unique.

(5) *(The) lion is a fine animal.
(6) *(The) winner is always happy.
(7) a. ?I met the student from our class at the supermarket.
b. I met a student from our class at the supermarket.
(8) a. We found the wallet in front of our door.
b. We found a wallet in front of our door.

H1d: L2 learners of L1s will incorrectly analyse the definite article as a specificity marker leading to (i) omission of the definite article for non-specific but definite nominals (e.g. generics) and (ii) failure to discriminate definite from indefinite specific nominals in relation to familiarity/novelty and uniqueness presuppositions.

H1e: L2 learners with L1s lacking a definite determiner will produce both (i) and (ii) type errors.

H1f: L2 learners with L1s with a definite determiner will not produce type (ii) errors but may produce type (i) errors (depending on how generics are encoded in their L1).

2.2. The learning path for acquiring specificity

An important question is whether the various aspects of specificity and definiteness will be acquired simultaneously. We hypothesise that they will not be acquired at the same time in English and that the nature of the L1 will determine the learning curves.

H2: Acquisition of specificity in English will be affected by (i) relevant properties of L1 grammars and (ii) qualitative and quantitative aspects of the input. Specifically:
H2a: Speakers of L1s with determiners will acquire specificity-related aspects of English grammar earlier than speakers of languages with L1s lacking determiners altogether. Further, such speakers will not have difficulty distinguishing specific definites from specific indefinites.
H2b: For all speakers, errors for which there is unambiguous and frequent input will disappear earlier than errors for which unambiguous input is infrequent.

Note that one consequence of H2a&b is that, depending on L1, a specificity-related error may characterize a lower level of proficiency for some sets of speakers and a higher level for others.

2.3. Methodology

The general hypotheses H1&H2 will be evaluated against the results of a systematic empirical investigation of H1a-f. The more specific hypotheses H1a-f will be tested on the CLC. It is obvious that specificity relates to linguistic interpretation in context. There is no linguistic item that is inherently specific or non specific. It follows that such a feature cannot be part of the description of any lexical item or syntactic category (and, of course, it is not currently coded in the CLC or in any corpus). A database that allows searches on a word/lexical level only would not enable us to test these hypotheses, therefore. But a database that allows us to manipulate information on parse trees, as the RASP parser does, offers the significant advantage that the linguistic environments in which specificity is relevant can be mapped onto the grammatical relations underlying the syntactic representations. Consider the case of predicative uses of nominals in (3) repeated below as (9). There is no single lexical item that can guide a search for this type of example. However, a database with syntactic trees defining grammatical relations allows us to search for e.g. structures of the type [NP BE NP]. Such a search will filter out irrelevant items like "Jane is waiting for a doctor", "Bill is stupid", "Planes are flying" etc. while at the same time returning examples of predicative structures such as (10).

(9) Jane is doctor.
(10) a. The kids are soldiers.
b. Jane is good doctor.

This approach to recovering linguistic environments in which specificity is relevant to abstract structural representations will guide the investigation of H1a-f.

2.4. Outcomes

We expect to be able to test the majority of the Hypotheses H1&H2. Descriptive results will be reported in a comprehensive report at the end of the year. Depending on the clarity and robustness of findings, we will pursue submission of a paper to an appropriate

conference on L2 acquisition. We expect this material to form the basis for a more serious journal publication in the course of the second year (with additional data).

Subproject (1.3): Morpho-syntax of Verbs (PI, Teresa Parodi)

Project Summary

The goal of this subproject is to examine finiteness in the morphology and syntax of English verbs. There is a developmental aspect to this: when do learners acquire certain morphological properties of verbs and their syntactic properties in the clauses in which they occur? There is also a crosslinguistic dimension: given that languages differ in the specification of these properties, predictions can be formulated as to how learners with different first languages (L1s) are likely to develop.

Research Goals and Outcomes

3.1 Background: Properties of English verbs

English distinguishes the classes of main verbs (eg. walk, drive), modals (eg. will, can) and auxiliaries (have, be, do) both morphologically and syntactically.

3.1.1 Morphological properties

Main verbs, auxiliaries and modals differ from each other in whether

a) they inflect for person or not:

(1) I walk he walks
(2) I am walking he is walking he does walk every day
(3) I can walk * he cans walk

b) they inflect for tense or not:

(4) I walk I walked
(5) I am / was walking, he did walk all the way
(6) I must / ! *was able to* walk

c) they have both finite and non-finite forms:

(7) I *cycle* I expect *to cycle*
(8) I *have* arrived I expect *to have* arrived
(9) I *can* cycle
(10) *I expect *to *can/ be able* to cycle

3.1.2 Syntactic properties

a) In terms of syntactic distribution, main verbs, auxiliaries and modals differ in whether they co-occur, as illustrated in the following table.

		modals *will*	auxiliary *have, be*	auxiliary *do*	main verb *drive*
modals	*will*	-	+	-	+
auxiliary	*have, be*	-	+	+	+
auxiliary	*do*	-	+	-	+
main verb	*drive*	+	+	+	-

(11) I *will have arrived*
(12) I *have arrived*
(13) I *have been driving*
(14) I *did* drive home
(15) *Did* you *drive* home?
(16) *Can* you *drive*?
(17) **Do* you *can* drive?
(18) **Have* you *can* drive?
but
(19) *Can* he *have* left?

b) Main verbs, auxiliaries and modals display different word order patterns.
In declarative sentences auxiliaries and modals precede, while main verbs follow adverbs such as 'always' and the negation.

(20) a. I will always swim
b. I can always swim
c. *I swim always.

(21) a. I will not swim
b. I cannot swim
c. *I swim not

In interrogative sentences, modals and auxiliaries precede the subject, while main verbs can only follow.

(22) a. can you arrive in time?
b. have you ever arrived in time?
c. are you arriving in time?
d. did you arrive in time?

but

(23) a. *arrive you in time?
b. *do you can arrive in time?
c. *have you can arrive in time?

3.2 Hypotheses
The differences just mentioned give rise to crosslinguistic and developmental questions, of relevance for the general hypotheses on the morpho-syntax of the verb complex (cf. §1.1.7 above) and on its syntax (§1.1.11). More detailed and more explicit versions of these hypotheses can be given as follows:

H1: The way different L1s categorise the morphological classes of modals, auxiliaries and main verbs will have an influence in the acquisition of these verbs in English.

H1a: Speakers whose L1 does not distinguish these classes morphologically (eg Romance languages, German) will transfer this to English and treat modals as main verbs.
(i) Person and tense markings will be overextended to English modals.
(ii) Modals will appear in non-finite contexts.
(iii) Verbs will take wrong complements, eg 'I must to leave' (for 'I must leave'), 'I will must take the bus' (for 'I will have to take the bus').

H2: Speakers of languages which do not distinguish a syntactic class of modals (Romance languages, German) will transfer this property to English and treat modals syntactically as main verbs.

H2a: Speakers of languages which do not distinguish a morphological and syntactic class of modals, will allow 'do' and modals to co-occur in questions, thus, treating modals as main verbs. ('Does Kim can drive?')

H2b: (i) Errors will decline with increasing proficiency, from A1 - C2.
(ii) Within each level the error rate can vary according to the nature of the task.

H3: Speakers of languages which allow main verbs to raise **(front?)** (Romance languages, German, South Slavic languages) will transfer this property to English.

H3a: Main verbs will have the same distribution of modals and auxiliaries.
(i) Main verbs will precede negation ('Kim walks not') and adverbs ('Kim walks always')
(ii) Main verbs will precede the subject in questions ('Walks Kim?')

H3b: *Do*-support will be misanalysed.
(i) Do-support will not be used in negative utterances. ('Kim walks not' as in H3a (i))
(ii) If do-support is used, tense will be marked both on *do* and on the main verb. ('Kim did not came').
(iii) Speakers of languages which use interrogative particles (eg. South Slavic languages) will treat 'do' as an interrogative particle and allow 'do' to co-occur with modals and auxiliaries. ('Does Kim can drive?', 'Does Kim has arrived?')

Notice that the type of error will be related to the properties of the L1. The same surface string may have different sources depending on the properties of the different L1s. The string 'Does Kim can drive?' can indicate that the learner is treating modals as main verbs (H2a) or that s/he is interpreting 'does' as an interrogative particle (H3b)

These hypotheses draw on crosslinguistic differences. From a developmental perspective the following can be expected:

H4: (i) Errors will decline with increasing proficiency, from A1 - C2, with some errors being more persistent than others.
(ii) Within each level the error rate can vary according to the nature of the task.
(iii) The rate of improvement will depend on the properties of the L1.

The formal approach proposed here looks at acquisition in a rule-based and organic way, as opposed to a list of items to be learned and a list of errors. This approach makes it possible to capture the developmental process from A1 to C2 and to formulate expectations as to the order of acquisition. It also allows us to capture the differences among languages and to formulate expectations for speakers of different languages. It can also shed light on cases in which learners give evidence of strategies which do not follow in a straightforward way either from their L1 or from the target L2.

Research on the acquisition of finiteness along the lines described here can start with speakers of Spanish, as representatives of the Romance group, followed by German as representative of the Germanic one. At a later stage and depending on the availability of relevant data in the corpus, a South-Slavic language can be added later. Current tagging can most probably provide a starting point.

This line of analysis has an obvious link with other components of the project. It can provide findings for the pedagogical component with respect to input and syllabus design involving the morphological and syntactic properties of English verbs. It is also relevant for assessment since it will document the stages at which these properties are mastered and, most importantly, the developmental paths followed by speakers of different languages.

PROJECT 2

From Strings of Words to Cohesive Discourse: Functional Progression from B2 to C2 (PI, Dr Henriëtte Hendriks)

(a) Project Summary

This project is complementary to project 1 in that it examines linguistic forms at the discourse level (verbs, prepositions, temporal markers, referential means), whereas project 1 mainly concentrates on morphology, grammar and lexis up to the sentence level. This project will look at three of the five domains that traditionally are thought to be of importance for discourse (Klein and von Stutterheim): *space, time* and *person* (events, and modality being domains 4 and 5) over the coming three years. The project will give an overview of form-function mappings at all levels of proficiency, thereby showing how these mappings become more and more appropriate in the context of the English language, and how learners become more and more native-like as a result of acquiring them.

Introduction and theoretical background

Previous cross-linguistics analyses of discourse structure show that languages differ in the choice of a particular referential frame, the attribution and management of the categories *topic-focus*, as well as the amount and type of information specified. This was long believed to be a result of differences in style, but recently it has been shown that this is more closely related to linguistic means available in the different languages. Although it is not entirely clear yet if the differences originate at the conceptualisation stage of speaking (Carroll and von Stutterheim), or rather at the thinking-for-speaking stage (Slobin), it is obvious that they seriously influence discourse production across languages. Thus, although German and English are closely related languages, when referring to time, and although both languages allow us to express the endpoint of an event, Germans, when shown a cartoon of a train running into a station and asked "*What happened*?" will speak only when the train has arrived at the station and say "*Der Zug ist am Bahnhof angekommen*" (the train has arrived at the station) whereas English speakers, asked the same question, will start speaking while the cartoon is still ongoing and say "*The train is running along the track, it is entering the station, it has arrived at the station*". Von Stutterheim and Lambert explain this difference through the accessibility of phasal verbs and the progressive in English, both of which are less accessible (even though not absent) in German. On a non-linguistic level, these differences are reinforced, in that Germans, while watching the cartoon will keep on following the train and its distance to its destination, whereas the English will simply concentrate on the train (von Stutterheim and Nüse, 2002 /results from eye-tracking experiments). The above differences between English and German speakers are quite systematic, although not absolute.

On the basis of such contrasts identified for speakers of different (but closely related) languages, the question arises as to how learners proceed when structuring discourse in a second language. Do they rely on the principles of information organisation preferably used in their mother tongue, which should lead to discourse deviating from native discourse, or do they acquire new strategies of information organisation together with a new linguistic system, or can they be located between the two poles? In the latest studies

on this subject, is seems that even near-native speakers, while producing grammatically perfectly acceptable discourse, to some extent adhere to the information structure as found in their native tongue. The above mentioned deviations from the target are extremely hard to identify in learners, and form part of the stylistic oddity often perceived in discourse produced by such advanced speakers. They will form the object of study of this second project.

The functional approach that underlies this proposal therefore allows us to look at language acquisition in a meaningful way across all levels of proficiency in the EPP (A1 to C2). It can look at the first occurrence of forms, at the functions initially attached to that form, and at the growth of form function mappings towards a target language mapping. Some of the possible linguistic means that could be singled out in an initial research programme could be motion verbs; temporal reference (including tense markers and adverbial markers); the article system. All of these linguistic means contribute ultimately to properly organized discourse, and have several other more local (utterance level) functions.

Year 1: Reference to space

Reference to space is traditionally expressed by the use of verbs (motion or location), prepositions and spatial adverbials. Typologically, languages will express the main component of spatial information, the path, either in the verb (verb-framed languages), or in so-called satellites (satellite-framed languages). Examples of the first category of languages are French, Spanish. Examples of the second category are English, and any other Germanic languages. Finally, there are verbs for which the typological placement is not entirely clear (Slavic languages, Chinese), and these are therefore also called equipollent languages.

As well described by Slobin in a number of articles, the above typological distinction does not only influence information encoded in the verb, but it also effects overall encoding of spatial components such as manner, cause, path and location, in that speakers of verb-framed languages, for example, will overall encode less manner than speakers of satellite-framed languages. The typological difference therefore influences not only the size of the verbal lexicon, but also the distribution of spatial information over verbs, prepositions and adverbial phrases on the utterance level, and the distribution of spatial information in the discourse as a whole.

Previous studies have shown that L2 learners will acquire prepositions in a preset order, and that they will also acquire the verbal lexicon in a more or less predictable order, in that verbs expressing more specific information (either manner or path or manner of attachment) will be acquired later than verbs that are more or less neutral in spatial information (put, go, come, take). Furthermore, the organisation of information over various parts of speech on the utterance level is very much ingrained in the speaker, and reorganisation is therefore not straightforward. This is particularly difficult given that many times there are a series of possible encodings, but one more preferred encoding for the native speaker. Hence, everything can be said, but not everything is habitually said. Finally, it has been shown in earlier studies that at the discourse level L2 learners frequently do not acquire the target language organisation. To give some examples:

When shown a little guy who is crossing a street pushing a ball so that it rolls in front of him (cf. Hickmann & Hendriks, 2006; Hendriks & Hickmann, 2007), English native speakers will typically comment on that situation as in (a). French will most typically comment on the situation as in (b). English learners of French, then, tend to construct possible but odd-sounding utterances as given in (c). The fact that these utterances are perfectly grammatical makes it particularly hard for learners to "unlearn" them. On a macro-level, the sentence level differences result in French speakers often distributing information over a number of clauses, as in (d), whereas the English native speaker will typically reply in a one-phrase utterance as given in (a).

(a) Hopi rolls the ball across the street
(b) Hopi traverse la rue en poussant le ballon
Hopi crosses the street while pushing the ball
(c) ?Hopi pousse le ballon en traversant la rue
Hopi pushes the ball while crossing the street
?Hopi roule le ballon en traversant la rue
Hopi rolls the ball while crossing the street
(d) Il ya Hopi. Il a un ballon. Il traverse la rue en poussant le ballon. Le ballon roule jusqu'à l'autre côté de la rue.
There is Hopi. He has a ball. He crosses the street pushing the ball. The ball rolls all the way to the other side of the street.

Goals

The overall goal of this project is to identify discourse structures in the target language and deviations from them in the learner languages. Such structures are rarely taught, and may be a result of the conceptualisation of a given task, or of the linguistic means available in a language. In any case, appropriating the correct discourse organization of a second language seems to be one of the last feature of that language many learners acquire. For the spatial sub-topic, I would propose to start with a Romance source-language, a Germanic source-language (German), and Chinese as a third language, it being classified as equipollent. This combination of languages will allow for language-specific hypotheses regarding source-language transfer. Once the analytic system can cope with three source-languages, it should be possible to quickly adopt it for all source-languages available in the corpus.

There are some obvious general searches of the database that are possible, just looking at the total inventory of verbs, prepositions and other spatial markers. This type of initial analysis would allow me to test order of acquisition hypotheses such as H1 and H2 below.

H1: Development in the verbal lexicon will progress from 1st to 3rd tier verbs (Slobin) (*put, go, come*, *leave, walk, jump,* and *skip, hop, pierce*) but source language type will influence speed of development.
H2: The prepositional lexicon develops in a universal order which is the same for L1 and L2 acquisition.

This will only be a first step in the analysis, however. The next step will involve making an inventory of the overall spatial information encoded (focus of information), and where in the utterance it is encoded (locus of information). This will specifically test hypothesis H3 and H4.

H3: In encoding spatial information on the sentence level, speakers will be less informative at lower levels of proficiency, and more informative (detailed) at higher levels of proficiency.

H4: Speakers will follow the L1 locus and focus of information in verbs and other means at lower levels of proficiency and only slowly adapt to the distribution of information as custom in the L2. This will result in different developmental paths, depending on source language.

Finally, one can look at the discourse organisation and test to what extent it is influenced by markings on utterance level. This will test hypothesis H5 as spelled out below.

H5: Only at very advanced proficiency levels will discourse organization start to resemble native speaker discourse organization as regards spatial information.

Outcomes

- a concrete list of features (on sentence and discourse level) that produce criteria to distinguish the various proficiency levels studied
- a list of features (sentence and discourse) that will distinguish profiles of learners of different L1s
- suggestions for source-language specific applications in teaching English as a second language
- at least one research paper prepared for submission in a scientific journal

References

Carroll, M., & von Stutterheim, C. (1993) The representation of spatial configurations in English and German and the grammatical structure of locative and anaphoric expressions. In: *Linguistics*, 31,6, pp.1011-1041.

Hickmann, M., & Hendriks, H. (2006) 1Static and dynamic location in French and English. *First Language*, vol. 26, 1, pp. 103-135.

Hendriks, H., & Hickmann, M. 2(2007) 'Caused motion in French by native speakers and English L2 learners' (Keynote address). To be presented at the international conference *Language learning and teaching in multilingual and multicultural contexts*, American University of Paris, Paris, France, October.

Klein, W., & von Stutterheim, C. (1987) Quaestio und referentielle Bewegung in Erzählungen. *Linguistische Berichte*, 109, pp. 163-183.

Slobin, D.I., (2003) Language and thought online: cognitive consequences of linguistic relativity. In: Deirdre Gentner and Susan Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought*, pp. 157-191. Cambridge, Mass.: MIT Press.

Von Stutterheim, C., & Lambert, M. (2005) Cross-linguistic analysis of temporal perspectives in text production. In: H. Hendriks (Ed.) *The Structure of Learner Varieties*, pp.203-230. Berlin: Mouton de Gruyter.

Von Stutterheim, C., & Nüse, R. (2002) Processes of conceptualization in language production: language-specific perspectives and event construal. In: *Linguistics*, 41, 5, pp. 851-881.

PROJECT 3
Visibility and Dissemination of EPP-related Research Findings through Conferences, Workshops and Websites

Prof John Hawkins, Dr Henriëtte Hendriks,
Dr Teresa Parodi (co-PIs):

(a) Project Summary, Goals and Outcomes

A key component of the English Profile Project is to disseminate the results of its research, to engage with those who test English worldwide using the CEFR, and more generally to be visible nationally and internationally within the fields of Applied Linguistics, Language Testing and Teaching, Linguistics, Psycholinguistics and Corpus Linguistics. It is important for Cambridge Assessment and for Cambridge University Press not only to incorporate the results of the English Profile Project in its testing services and publishing, but also to be seen to be a leader worldwide in these research areas. Dissemination requires a regularly updated website, presentations by project members at major conferences in these fields, workshops specifically for EPP participants and stakeholders, and high-profile and high-prestige academic events organised by project members.

1 EPP Website

The English Profile website represents an important means of interacting with stakeholders. At the moment it consists of a largely static website containing introductory information regarding the project and a questionnaire for teachers, plus a members' only area containing important documents to be shared by the various groups involved. Caroline will leverage her expertise in web design and her understanding of the concerns and priorities of researchers in language acquisition to create a more dynamic website which engages current and potential project researchers and helps to promote stakeholder involvement.

2 Presentations by RCEAL members at major conferences

- John Hawkins will give a keynote presentation at the ALTE conference, Spring 2008, describing the Corpus Linguistics Project and other UCLES-RCEAL projects within the English Profile Project

- Paula Buttery, Caroline Williams and other project members will lead break-out sessions at ALTE, following up on information presented in JH's plenary address

- Paula Buttery LSA Meeting - The Linguistic Society of America have agreed to allow Paula to raise awareness of the English Profile Project by addressing attendees of the Linguistic Institute 2007---"Empirical Foundations for Theories of Language". This year's Linguistic Institute will be held in Stanford in July and has attracted a world-wide and world-class audience of students and academic staff. This is the perfect platform for disseminating information about the EPP in

the USA and it should set the stage for attracting participants to our language acquisition workshop and conference (see below) among a broad range of linguists with strongly interdisciplinary interests.

- Paula Buttery and Caroline Williams will work towards presenting publications at BAAL 2008 in Edinburgh.

3 Workshops for EPP participants and stakeholders

- February 2008 EPP workshop for local EPP participants and stakeholders
- Other local events.

4 High-prestige and high-visibility academic events

The testing of EPP hypotheses and predictions to date (cf. §§1.2.1-4 in project 1) has uncovered some issues that get to the very heart of current theories of language acquisition. The six levels of the CEFR are based ultimately on the practical experience of examiners, who find it feasible and useful to assign learners to these different levels (and to finer levels of attainment within each CEFR level). The theoretical and descriptive work of the EPP is designed to examine developmental stages in second language acquisition, in order to better understand the basis for examiners' judgments and in order to contribute to improved testing and validation in the future. It is therefore important that we share our EPP findings and research issues with the international research community. To this end RCEAL is planning two events over a 2-year period: a workshop on the topic of "Developmental Stages and Learner Profiling" which will bring together experts from different areas of language acquisition with whom we can discuss foundational issues and from whom we can learn; and a major international conference on this same theme one year later, for which the workshop will serve as preparation.

Workshop Description: Developmental Stages and Learner Profiling

From the earliest studies on language acquisition, there has been a strong desire to establish an index of language development. This has been especially true for second language acquisition, but even in first language acquisition many studies have tried to establish specific orders of acquisition and their causal factors. N. Ellis and Larsen-Freeman, in 2006, in the special issue of Applied Linguistics on Language Emergence comment that “such an index would be a ‘boon’ as it would eliminate vagaries associated with classifying learners as beginning, intermediate, and advanced” (p. 564). The more we study acquisition, however, the more we realize that describing developmental stages is as complex as describing language itself. It is rare for a given linguistic phenomenon to be uniquely associated with a certain developmental stage. Typically each stage is characterized by a cluster of phenomena. It has also become clear that some indicators may work at a specific moment in development, but not at another point in time. The properties of the languages in question will also have an impact on the acquisition process and as a result we see similarities and differences in developmental paths, both in L1 and in the L2 acquisition of different languages. In early bilingual and in L2 acquisition, there is the additional challenge that two languages interact. It has also become clear that language development is by no means linear, rather it seems to progress in recurring waves, involving more static periods interspersed with big leaps forward. Finally, and most significantly, it has become clear that even when we think we have found a cluster of phenomena that can indicate a given stage in language

development, this clustering may hold in different ways for different groups of learners, due to the factors mentioned above.
In this workshop we propose to bring together researchers who have tried to establish such developmental sequences. They may have based their index on more formal properties of language or on more communicative or functional properties. Specifically, we propose to bring together researchers who have worked with five different populations of learners: those who work with normally developing monolingual first language learners; those who work with normally developing bilingual language learners; those who work with atypical developing first language learners (SLI, autism, deaf or blind language learners); those who work with non-guided child and adult L2 learners; and those who work with guided (i.e. taught) child and adult L2 learners.

Given the multifaceted character of language and of language learners, our hope is that bringing such researchers together will allow for a much-needed cross-fertilization of ideas regarding research methods, causal factors, linguistic methodology, etc. The format of the conference, and the selection of invitees, will be designed to encourage such cross-fertilization.

PROJECT 4

Compiling New Corpus Data and Rendering Them Searchable for Criterial Feature Identification A2-C2

Dr Dora Alexopoulou, PI

(a) Project Summary

There is a consensus within all research groups of the English Profile Project that more data should be obtained in order to extend the current database. There is also a shared view that new data should go beyond the current exam scripts. As a result, plans are being drawn up for collection of new data primarily through contacts with British Council centers worldwide. The primary aim of this project is to contribute to a comprehensive plan for the collection of data and for the extension of the current database. Close consultation with other research teams within EPP will be maintained throughout.

Research Goals and Outcomes

The aim of the project is to offer advice on the following issues with an eye to identifying and prioritizing the kinds of data that should be collected.

(i) **Learners**:

a. Which learning stages should be prioritized? Currently there is a scarcity of data for the lower levels; on the other hand, despite larger samples for the advanced levels, the criterial features distinguishing advanced levels are less well understood.
b. Which L1s should be prioritized? Should we give priority to languages for which there is already a good sample so as to be able to perform more reliable statistical analysis, or should we give priority to languages for which there is currently less data available, so as to bring all languages to the same level?
c. Other demographic features of learners: age (at the time of the test, age at which English lessons started), education (private/state), education for English language (private/state), motivation (why learn English) etc. might be relevant factors.

(ii) **Types of materials** to be collected:

a. Discourse genre: different discourse genres (composition, letter, journal, dialogue, e-mail) involve different linguistic competencies and can reveal different types of structures and vocabulary. Should one be prioritized over the other? Is it possible to aim for a balanced mix within the time and other practical limitations we'll be operating with?
b. Topics: as with discourse genres, different discourse topics can reveal varieties of linguistic competencies. Easy to talk about topics might yield more text, hence more data, while topics requiring a more structured argument might yield less text but a more interesting set of linguistic structures. Choice of topics may vary depending on the level of the learner.

c. Types of errors: are there specific types of errors of particular interest? If yes, are they are more likely to occur in a particular type of text or in a particular type of topic?

(iii) **Input**:

a. Teachers: Would it be useful to collect information on the teachers and teaching strategies in relation to the following? 1) demographic features of teachers (e.g. whether they are native speakers of English or second language learners, if the latter where they learnt English, why they became English teachers, education); 2) who teaches what (e.g. in Greece Greek teachers would teach the grammar part of the exam and native speakers the reading, comprehension, and conversational parts of the exam); 3) should we collect data from teachers as well?
b. Teaching materials: what kinds of textbooks are used? Which kind of "free" materials (i.e. collected through newspapers, magazines etc by teachers) are used?
c. Intensity of input: frequency of classes, availability of English on TV, cinema etc.

(iv) **Method of collection**:

a. What is the best way to obtain materials in an electronic form as soon as possible?
b. How can the EPP website facilitate the collection process?
c. Would it be possible to set up "online chats" with native speakers in Britain (say school children volunteering in Cambridge schools) through the website to obtain naturalistic dialogue data?

The PI will seek advice from other members of the Corpus Linguistics Group within RCEAL and will be in regular consultation with other research groups within EPP for addressing issues i-iv and, others, as they arise in the course of these consultations.

The main outcome of this project will be a comprehensive document advising on issues i-iv.