Communicative Adequacy and Linguistic Complexity in L2 writing

Folkert Kuiken, Ineke Vedder and Roger Gilabert
University of Amsterdam / University of Amsterdam /
University of Barcelona

The chapter investigates the relationship between communicative adequacy and linguistic complexity (syntactic complexity, lexical diversity, accuracy) of the written output of L2 writers of Dutch, Italian and Spanish. The main goal of the *CALC* study ('Communicative Adequacy and Linguistic Complexity') discussed in the chapter is to investigate the relationship between the communicative aspects of L2 writing, as defined in the descriptor scales of the Common European Framework of References (CEFR, Council of Europe 2001), and the linguistic complexity of L2 performance. It is argued that the interpretation of syntactic complexity, lexical diversity and accuracy is not possible without also taking into account the communicative dimension of L2 production.[1]

1. Introduction

The notion of language proficiency presented in the Common European Framework (CEFR) rests on two pillars, as has been pointed out in several studies (Hulstijn, 2007). Language proficiency is defined both functionally ('can-do statements'), describing the number of domains, functions and roles language users can deal with in the L2 (*what*), and in terms of the quality of language proficiency, e.g. the degree to which language use is effective, precise and efficient (*how* well; Hulstijn, this volume). Whereas the majority of research conducted so far has been concerned with the can-do-statements and the functional scales of the CEFR (Little, 2007), fewer studies have focused on the linguistic dimension, particularly regarding the question of whether it is possible for L2 learners to be situated at different linguistic scales and levels (for instance the

[1] We would like to thank the raters for their evaluation of the data of Dutch, Italian and Spanish for both the L2 writers and the L1 writers. We also thank Luuk Nijman for his invaluable help with the statistical analysis of the data. Finally we thank the two anonymous reviewers for their useful comments on an earlier draft of this chapter.

B1 level for vocabulary range, and the A2 level for grammatical accuracy), or the specific ways in which L2 proficiency develops in different European languages. Moreover, the CEFR doesn't indicate, for a given target language, which particular developmental features can be identified as being characteristic for a given scale level (Alderson, 2007). The relationship between language proficiency and language acquisition and the overall development of L2 proficiency (in terms of syntactic complexity, lexical diversity, fluency and accuracy) and the way in which they interact, is thus still unclear (Hulstijn, 2007, this volume).

The relationship between the functional descriptor scales of the CEFR on the one hand and the linguistic scales on the other hand has not been addressed much in the literature either. One of the few studies which have investigated the relationship between the functional and the linguistic dimension of L2 performance is the so called *WISP* study ('What Is Speaking Proficiency'; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2007, in press). In the *WISP* study the oral performance of 208 L2 speakers and 59 native speakers of Dutch was examined both in terms of communicative success and in linguistic terms, concerning the mastery of a number of linguistic skills, such as fluency (i.e. breakdown fluency, speed fluency and repair fluency), syntactic complexity, and vocabulary control. The main question in this type of research is to what extent it may be expected that L2 learners who are situated at the B2 level of the functional descriptor scales of the CEFR have also attained the B2 level with regard to their linguistic performance. In other words, the issue at stake is if and how the communicative adequacy of L2 performance ('getting the message through') is related to the syntactic complexity, lexical variation, fluency, and accuracy of the output.

Whereas in several studies on L2 speaking and writing general measures for assessing the complexity, accuracy and fluency (CAF) of L2 performance are employed, few studies in SLA research report on the communicative success and adequacy of the L2 output. This, however, is in clear contrast with language teaching practice and testing, where often both the communicative dimension and the linguistic complexity and accuracy of the L2 production are independently assessed (Pallotti, 2009). A possible reason for the paucity of studies which explore the relationship between the communicative adequacy of the L2 production and the linguistic forms by means of which the message is conveyed may be due to the absence in the literature of a coherent and clear-cut definition of communicative adequacy as a construct. While communicative adequacy is often interpreted as socio-pragmatic appropriateness (McNamara & Roever, 2007), in other cases it is mainly considered in terms of communicative effectiveness (i.e. success of information transfer; Upshur & Turner, 1995) or successful task completion (i.e. relevance and effectiveness of content according to task instruction; De Jong et al., 2007, in press; Pallotti, 2009). In the present chapter communicative adequacy is interpreted as a task-related, dynamic and interpersonal construct, focusing both on the specific communicative task which has to be carried out by the speaker or writer (e.g. writing an email to a friend to suggest a restaurant for dinner), and the way the message is received by the interlocutor (the listener or reader).

There is no unanimity in the literature either as to how communicative adequacy could best be assessed. Contrary to CAF measures, general and quantitative measures to rate communicative adequacy are lacking. Moreover, it is not clear by which textual and linguistic features communicative adequacy, in the eyes of raters, is mainly determined (see however the study by Iwashita, Brown, McNamara, & O'Hagan, 2008, for an investigation of the relationship between certain features of the oral L2 production of test-takers and the holistic scores awarded by raters to these performances). In order to be able to assess communicative adequacy, it is thus necessary to resort to proficiency scales, like the ones of the CEFR, containing a set of descriptors to evaluate the features of L2 performance relevant for a particular level of proficiency. However, as has been pointed out in a number of studies, one of the problems of the use of proficiency scales is that they are generally not calibrated or empirically validated and that they do not refer to any theoretical paradigm (see also the chapters in this volume by Alanen, Huhta, & Tarnanen; Hulstijn, Alderson, & Schoonen; Pallotti).

In order to explore the role of communicative adequacy in L2 writing and to establish whether, at a given level of L2 proficiency, communicative adequacy and linguistic complexity develop at the same pace or at the expense of each other, the *CALC* study ('Communicative Adequacy and Linguistic Complexity') was set up. The basic assumption, underlying *CALC*, is that syntactic complexity, lexical diversity, and accuracy cannot satisfactorily be interpreted without taking into account the communicative adequacy of the L2 text. The *CALC* study examines the extent to which the communicative adequacy of the written L2 production is related to the linguistic complexity and accuracy of the text. The corpus on which the analyses have been conducted consists of 206 short written essays. Participants in the study are 34 L2 learners of Dutch, 42 L2 learners of Italian, and 27 L2 learners of Spanish. To create a baseline comparison, the writing tasks have also been administered to a group of 18 native speakers of Dutch, 22 native speakers of Italian, and 10 native speakers of Spanish. In this chapter the data of the L2 learners of Dutch, Italian and Spanish are discussed.

The main goal of CALC is to provide evidence of learner performance, both in communicative and in linguistic terms (i.e. grammar, lexis, accuracy), at a particular scale level of the CEFR. More specifically, the study investigates the relationship between the communicative adequacy and the linguistic complexity, operationalized as syntactic complexity, lexical diversity and accuracy, of learner output elicited by two writing tasks at the B1 level of overall written production of the CEFR, e.g. a short essay on a topic of interest for a particular functional purpose, in which an opinion has to be reported about factual information (Council of Europe, 2001).

A second aim of the CALC project is to contribute to the description of interlanguage and the role of proficiency in L2 writing by analysing the use of particular linguistic features and structures that typically characterize L2 performance at a given proficiency level, such as the elaboration of the noun phrase and the use of subordinate clauses. Finally, the study investigates the learning dimension of L2, in relation to the CEFR levels. The outcomes of the study are thus relevant for assessment and syllabus design. In this chapter we therefore focus on the first goal of the CALC study.

2. CALC: Design of the study

2.1. Research goals

The main goal of the present study, as pointed out above, is to investigate the relationship between communicative adequacy and linguistic complexity in L2 writing. In more general terms and related to this main goal, this study also aims at contributing new data to the description of interlanguage by using 'diagnostic' linguistic measures which may shed light on the role of L2 proficiency in writing. In order to achieve such goals, students with three different target languages were asked to perform two tasks in writing, and their productions were both rated holistically and measured by means of standardized measures of L2 writing performance. The level of proficiency of the students ranged from A2 to C1, although a large majority of them fell within the range A2-B1.

2.2. Research questions

The following questions were formulated in relation to the goals of this study:

1) What is the relationship in L2 between communicative adequacy as assessed by individual raters, and linguistic complexity (i.e., syntactic complexity, lexical diversity, and accuracy), as assessed also by the same individual raters?
2) What is the relationship in L2 between communicative adequacy, as assessed by individual raters, and linguistic complexity, as assessed by general measures of linguistic complexity?
3) What is the relationship in L2 between linguistic complexity, as assessed by individual raters, and linguistic complexity, as assessed by general measures of linguistic complexity?

The first research question aims at exploring whether a correlation exists between L2 learners' communicative adequacy when performing the task and their linguistic performance. In order to answer this question, holistic measures based on the CEFR descriptors are used and written productions are assessed by individual raters. The second research question tackles the issue of whether communicative adequacy as holistically rated by experienced[2] raters correlates with linguistic complexity, which is calculated this time by means of general linguistic measures (i.e., measures of structural complexity, lexical diversity, and accuracy, which are further described below). The third question deals with the potential correlation between linguistic complexity as perceived by individual raters using holistic measures and linguistic complexity as analysed and calculated by means of general measures of linguistic complexity.

Given the paucity of studies in this area[3] to motivate any directional hypothesis, no specific hypotheses are advanced. We have no sufficient grounds to hypothesize whether communicative adequacy will develop at the same pace as or separately from linguistic complexity. Our study is thus what Seliger & Shohamy (1989, p. 29) have labelled as heuristic or hypothesis generating kind of research.

2.3. Participants

Three groups of university students participated in the study. One group consisted of 34 international students learning Dutch as a second language. Their average age was 26,1 years. There was a wide variety of L1s in this group. Another group consisted of 42 Dutch students who had Italian as a foreign language. Their average age was 21,5 years. The third group consisted of 27 Dutch students taking Spanish as a foreign language. Their average age was 24.9 years. All of the students were enrolled in the modern language section of the University of Amsterdam.

2.4. Materials

Two communicative tasks were used in this study (see Appendix 1 for an example of task 1). Both tasks were similar in terms of type and structure. In both tasks learners were required to make a decision about which of three non-gov-

[2] By 'experienced' we mean language teachers who have experience testing students orally, and so may be able to subjectively assess whether a learner is doing a good job communicating a message or not.

[3] Although not specifically dealing with the issue of 'communicative adequacy', see Iwashita et al. (2008) for a large scale study on the issue).

ernmental organizations to choose as a candidate for receiving a grant in task 1, and which of three topics presented to the learners they would like to see published in their favourite newspapers in task 2. Both tasks were open in the sense that learners could choose from a number of possibilities. The communicative goal of the tasks was to provide arguments to convince a university board in task 1 and a board of journalists in task 2 to choose their recommended options. Learners were given four instructions in each case: to specify which organization and topic they would support; to describe the aims of the organization and the importance of the topic; to indicate the beneficiaries of the organization's work and the readers that would potentially be interested in the article of their choice; and to provide at least three reasons to convince their addressees. The two tasks were designed with CEFR descriptors which are associated with B1 level, and therefore accessible for Dutch L2, Italian L2 and Spanish L2 participants in our study. Students were told they had 35 minutes for each task and they were told to write at least 150 words (i.e., roughly 15 lines).

2.5. Procedures

Data collection took place during a two-week period. Learners were contacted in class and they were briefly told about the research project. All students participated on a volunteer basis. They were informed that the projects would help researchers understand what students are able to do at each CEFR level. Then learners took a C-test[4], which is described below, and they were also asked to fill out a personal data questionnaire which was either administered during this session or at the teacher's discretion. After having completed the C-test half of the learners were presented with task 1, while the other half started with task 2 and vice versa. (see Appendix 1, example of task 1). Immediately after task performance the students were asked to fill out a perception questionnaire which asked them about the difficulty in performing the task, their own evaluation of their performance, and the interest of the task. Students performed the second task under the same circumstances as the first task. Again, after finishing the second task, they had to fill out a perception questionnaire.

[4] DIALANG was used as a backup proficiency test but is not reported here. The DIALANG is the test associated with the CEFR, and it provides an indication of the current level of the test-taker at a given point. In this test, which is a subset of the whole DIALANG test, learners are asked to look at lists of verbs in the target language and decide whether they are verbs that exist in the target language. The test provides the learner with a score and places him or her within one of the CEFR levels. In this study only C-test proficiency results are reported.

2.6. Instruments and measures

A number of instruments and measures were used to calculate the learners' proficiency, their productions in holistic terms, and the linguistic dimensions of their written productions. Regarding proficiency, a C-test was used in which learners are asked to complete 100 words in five short texts in which half the letters of every other word have been replaced by blanks. Learners are asked to reconstruct the words by considering contextual clues. The C-test has been shown to correlate with other general proficiency tests (see Babaii & Ansary, 2001; Jafarpur, 1999; Klein-Braley, 1997). Beyond their already assessed discriminatory power and standardized use in the literature, the criterion used to select this test was the fact that it could be completed in just 15 to 25 minutes. Given that data were collected in a classroom context it was important for the tests not to take too long and not to disrupt the class too much.

For the holistic rating of learners' productions, the researchers drew on two main sources: the general descriptors provided by the CEFR on the one hand and the measures developed for the calculation of speaking proficiency by the *WISP* group at the University of Amsterdam, which were also inspired by the CEFR, on the other hand. Based on these two sources, the criteria used for holistic rating (See Appendices 2 and 3) were adapted to the specific tasks learners were presented with. Meetings with the raters for each target language were held in which holistic assessment was presented, discussed, and piloted on a small number of productions. When sufficient agreement was reached, raters were given written productions which were assessed by means of the holistic criteria of communicative adequacy and linguistic complexity. They were asked to do this on their own time and were instructed to rate the productions separately for communicative adequacy and linguistic complexity. For Dutch four raters were asked to judge each text, for Italian and Spanish there were three raters.

As for the general linguistic measures, standardized measures in both oral and written task performance literature were used.

Syntactic complexity:	Clauses per T-unit Subclause ratio
Lexical diversity:	Guiraud Index of Lexical Richness
Accuracy:	Total number of errors per 100 words Total number of errors per T-unit

Clauses per T-Unit and subclause ratio are two well-established measures of structural complexity (Wolfe-Quintero, Inagaki, & Kim, 1998). For lexical diversity the Guiraud Index of Lexical Richness (see Vermeer, 2000, for an eval-

Table 1. Interrater reliability scores as assessed by Cronbach's alpha

L2	Task 1		Task 2	
	Comm. Adeq.	Ling. Compl.	Comm. Adeq.	Ling. Compl.
Dutch L2	0.761	0.889	0.777	0.882
Italian L2	0.702	0.868	0.752	0.735
Spanish L2	0.700	0.756	0.717	0.793

Table 2. Means and standard deviations of measures used in the experiment

	Means and *Standard Deviations*															
	Clauses/ T-unit		Dep.C/ Clause		Guiraud		Tot.Err/ T-unit		Tot.Err/ 100 w.		C-test		Comm. Adeq.		Ling. Compl.	
Dutch L2																
Task 1	1.94	*.83*	.46	*.12*	7.00	*.62*	2.44	*.22*	17.95	*8.06*	78.31	*8.13*	2.99	*.84*	2.82	*01*
Task 2	1.84	*.30*	.45	*.09*	6.88	*2.73*	2.51	*1.19*	18.98	*8.37*	77.29	*7.70*	3.11	*.85*	2.59	*.97*
Italian L2																
Task 1	1.87	*.35*	.45	*.10*	6.47	*1.11*	1.88	*.90*	15.61	*6.83*	69.49	*12.81*	2.97	*.86*	2.19	*.93*
Task 2	1.94	*.41*	.47	*.11*	6.57	*.94*	1.85	*1.30*	15.58	*9.46*	69.43	*12.41*	2.98	*.84*	2.39	*.80*
Spanish L2																
Task 1	1.89	*.27*	.46	*.07*	6.87	*.74*	1.06	*.60*	6.18	*3.45*	83.54	*.55*	2.39	*.71*	2.50	*.81*
Task 2	1.80	*.22*	.44	*.22*	6.87	*.81*	1.20	*.73*	7.20	*4.08*	83.08	*8.54*	2.35	*.80*	2.60	*.89*

uation of the measure) has shown to discriminate among learners at different levels. One of its advantages is that it corrects for differences in text length. In order to discriminate among the different levels of accuracy, two standardized measures in the psycholinguistic and the task-based performance literature were used. The total number of errors per 100 words and the total number of errors per T-units also compensate for differences in text length.

2.7. Statistical instruments

Both descriptive statistics and correlations are used in this study. Descriptive statistics are used to specify means and standard deviations, and Pearson correlations are applied to capture the potential relationship between communicative adequacy and linguistic complexity as measured by raters, and by these two holistic measures and the general measures of performance employed in the study. As will be seen below, correlations were calculated separately for task 1 and task 2. Cronbach's alpha was used for the calculation of interrater reliability.

3. Results

First, interrater reliability for Dutch L2, Italian L2 and Spanish L2 was assessed by means of Cronbach's Alpha, both for tasks 1 and 2 and for communicative adequacy and linguistic complexity (for results see Table 1). The interrater reliability coefficients can be considered sufficient to good, as they varied from 0.700 (Spanish L2, task 1, communicative adequacy) to 0.882 (Dutch L2, task 2, linguistic complexity). In general interrater reliability scores tend to be higher on linguistic complexity than on communicative adequacy.

The descriptives (i.e. means and standard deviations) of the measures that have been used in order to answer our research questions are presented in Table 2. At first sight these numbers look rather stable over the two tasks and the different measures used for the three groups of L2 learners. Perhaps the only salient finding is that the Spanish L2 learners make fewer mistakes than the Dutch L2 and Italian L2 learners. They also obtain higher scores on the C-test, so it might be the case that their general level of language proficiency is higher. However, this is not reflected in the scores of the raters on communicative adequacy and linguistic complexity.

We also considered a possible interdependency of the measures used: for instance, if T-units are longer, then there is more room for errors to be made. Using Pearson correlations we found a positive correlation between the number of clauses per T-unit and the number of errors per T-unit for Dutch L2-learners on task 1 ($r = .366$, $p < .05$), and for Spanish L2 learners on task 1 ($r = .505$, $p < .01$); for Spanish L2 learners on task 1 the correlation between the number of dependent clauses per clause and the total number of errors per T-unit was significant as well (r = .504; p < .01). On the other hand, for Italian L2 learners on task 1, the syntactic measures correlated significantly with the Guiraud index (number of clauses per T-unit: r = .364, p < .05); number of dependent clauses per clause: r. = .365, p < . 05). For task 1 we also noted a (negative) correlation between Guiraud and the number of errors per 100 words (r = -.340; p < .05).

In order to answer the three research questions regarding the relationship between communicative adequacy and linguistic complexity Pearson correlation coefficients were calculated. We looked at the correlation between these two variables in two ways: bivariately and controlling for the participant's proficiency, as measured by their score on the C-test.

Our first research question concerns the relationship between communicative adequacy and linguistic complexity, both assessed by the raters on a six point Likert scale (see Table 3). Pearson correlation coefficients varied bivariately from 0.604 (Spanish, task 1) to 0.827 (Italian, task 1) and can be considered moderate to good. Taking into account the proficiency level of the participants, the correlation coefficients decreased (from 0.479 for Spanish on task 2 to 0.653 for Italian on task 2). Nevertheless, all correlations but one remained significant at p < 0.01.

Table 3. Pearson correlations between communicative adequacy and linguistic complexity, both based on ratings on a six point Likert scale

L2	Bivariate		Controlling for proficiency	
	Task 1	Task 2	Task 1	Task 2
Dutch L2	0.820**	0.768**	0.650**	0.559**
Italian L2	0.827**	0.777**	0.639**	0.653**
Spanish L2	0.604**	0.636**	0.534**	0.479*

* p < 0.05; ** p < 0.01

Our second research question regards the correlation between communicative adequacy as assessed by individual raters on a six point Likert scale and linguistic complexity as assessed by general measures of syntactic complexity, lexical diversity and accuracy. As mentioned in section 2, the measures we used were the number of clauses per T-unit and the number of dependent clauses per clause for syntactic complexity, the Guiraud index for lexical diversity, and the number of total errors per T-unit as well as the number of total errors per 100 words for accuracy (for results see Table 4). If we first consider the bivariate correlations in Table 4, we notice that no significant correlations can be established for the measures of syntactic complexity (clauses per T-unit, dependent clauses per clause), whereas almost all correlations for lexical diversity (Guiraud) are significant (except for Spanish on task 1); the same holds for accuracy (total errors per T-unit, total errors per 100 words), since all correlations are significant except for Italian on task 1. However, if we calculate these correlations by factoring in the proficiency level of the participants, the correlation coefficients radically drop and the significant correlations decrease in number. All in all, this seems to indicate that raters, when making communicative adequacy judgments, rely more on lexical diversity and accuracy than on syntactic complexity.

Table 4. Pearson correlations between communicative adequacy as assessed by raters on a six point Likert scale and linguistic complexity as assessed by general measures

CORRELA-	Bivariate					Controlling for proficiency				
TIONS	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w
Dutch L2										
Task 1	-,015	,079	,582**	-,541**	-,676**	-,353	-,241	,305	-,531**	-,495**
Task 2	,090	,100	,376*	-,394*	-,531**	,010	,010	,278	-,322	-,279
Italian L2										
Task 1	,288	,251	,671**	-,199	-,473**	,105	,088	,318*	,080	-,064
Task 2	-,066	-,016	,568**	-,453**	-,578**	-,226	-,195	,320	-,348*	-,318
Spanish L2										
Task 1	-,279	-,254	,262	-,732**	-,713**	-,361	-,345	-,034	-,717**	-,670
Task 2	-,059	-,310	,636**	-,638**	-,580**	-,189	-,159	,543	-,504*	-,515

* $p < 0.05$; ** $p < 0.01$

A similar picture emerges if we turn to the third research question, concerning the relationship between linguistic complexity as assessed by the raters on a six point Likert scale and linguistic complexity as assessed by the general measures mentioned above (for results see Table 5). Again, concentrating first on the bivariate correlations, Table 4 shows that there are no significant correlations for the measures in terms of syntactic complexity, whereas almost all correlations for lexical diversity are significant (except for Dutch on task 2 and Spanish on task 1), while all correlations are significant with respect to accuracy. Also, taking into account here the proficiency level of the participants, the correlation coefficients tend to drop and the number of significant correlations decreases, although less drastically than in Table 4: most of the correlations concerning accuracy stay significant (except for Italian on task 1), while the correlations on lexical diversity also remain significant for Italian. As with respect to raters' judgements on communicative adequacy, raters also seem to rely more on lexical diversity and accuracy when making linguistic complexity judgments.

Table 5. Pearson correlations between linguistic complexity as assessed by raters on a six point Likert scale and linguistic complexity as assessed by general measures

CORRELA-	Bivariate					Controlling for proficiency				
TIONS	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w	Clauses/ T-unit	Dep.C/ Clause	Guiraud	Tot.Err/ T-unit	Tot.Err/ 100 w
Dutch L2										
Task 1	,049	,075	,433*	-,757**	-,873**	-,197	-, 194	,081	-,824**	-,833**
Task 2	-,200	-,212	,197	-,726**	-,816**	-,352	-,292	,044	-,756**	-,738**
Italian L2										
Task 1	,213	,188	,673**	-,352*	-,566**	-,015	-,012	,313*	-,157	-,229
Task 2	-,038	,025	,634**	-,471**	-,584**	-,224	-,177	,378*	-,366*	-,334*
Spanish L2										
Task 1	-,249	-,226	,173	-,725**	-,777**	-,349	-,317	-,150	-,660**	-,689**
Task 2	,188	,194	,398*	-,641**	-,761**	,048	,050	,216	-,491*	-,623**

* $p < 0.05$; ** $p < 0.01$

Because the participant's proficiency level as measured by the C-test seems to play a substantial role if it is being controlled for, it was decided to look into this effect more thoroughly. Therefore the groups were split up into two subgroups depending on their score on the C-test. Participants with a C-score in the lowest 40th percentile were assigned to the 'low level' subgroup, and students with a C-score ranking into the highest 40th percentile were placed in the 'high level' group. The intermediate category was excluded from this part of the analysis. Next, the correlations between communicative adequacy and linguistic complexity, both assessed by the raters on a six point Likert scale, were established for each subgroup separately (for results see Table 6). As can be seen from Table 6 these correlations turned out to be significant for Dutch L2 and Italian L2, but not for Spanish L2. It also appears that, in general, the correlations tend to be higher for the high level group in comparison to the low level group.

Table 6. Proficiency level (based on C-test) versus correlation between communicative adequacy and linguistic complexity on a six point Likert scale

L2	Task 1	Task 2
Dutch L2		
LowCtest	,758**	,678**
HighCtest	,854**	,807**
Italian L2		
LowCtest	,575*	,677**
HighCtest	,759**	,772**
Spanish L2		
LowCtest	,340	,615
HighCtest	,534	,225

* $p < 0.05$; ** $p < 0.01$

4. Conclusion and discussion

First of all, the reliability among the raters as measured by Cronbach's alpha was sufficient to good, but the reliability scores for linguistic complexity tended to be higher than for communicative adequacy. All correlations were significant, and they remained significant when we took into account the proficiency level of the participants. In our view, then, raters reached a reasonable level of agreement in the interpretation of both scales. Raters, however, seemed to agree more clearly on their interpretations of the linguistic complexity criteria. It may be the case that experienced raters may have more often dealt with linguistic criteria than with functional ones, which may explain some differences in the interpretation of the communicative adequacy criteria. It may also be the case that the way the various levels of language proficiency were defined in terms of linguistic complexity was more clear to the raters than in terms of communicative adequacy.

If the participants were split up according to their proficiency scores based on a C-test, it appeared that generally the correlations between communicative adequacy and linguistic complexity tended to be higher for the high level group than for the low level group. Our findings also suggest that communicative adequacy and linguistic complexity seem to be more balanced in the case of advanced learners. This is less the case for lower level learners who may concentrate either on communicative adequacy or on linguistic complexity, and for whom it is probably more difficult to focus on communicative adequacy while they are still struggling with form. Another explanation for the higher correlations of the more advanced learners might be that they tend to use longer sentences, which might encourages raters to give them a higher score on communicative adequacy.

With regard to the results of the first question, there are at least two possible explanations as to why high correlations between communicative adequacy and linguistic complexity were obtained, one which refers to the raters and one to the learners themselves. The first one is that the raters either may have perceived linguistically complex compositions as also communicatively adequate or vice versa. Such high correlations suggest that the development of communicative adequacy and linguistic complexity may go hand in hand. As pointed out by Alanen et al. (this volume) the accuracy and complexity of grammar and vocabulary will always have some influence on the communicative adequacy of the L2 production and the extent to which learners are able to complete the task they are rated on. Therefore the linguistic features will influence the ratings of the communicative adequacy. The second explanation is that more proficient learners, who obtained a higher score for linguistic complexity, may also have had more attentional and memory resources to deal with communicative adequacy, while lower level learners need to devote their cognitive resources to working out language problems, at the expense of the communicative and functional aspects of task performance.

The second research question concerned the correlation between communicative adequacy as assessed by individual raters on a six point Likert scale and linguistic complexity as assessed by general measures of syntactic complexity, lexical diversity, and accuracy. We found significant correlations for lexical variation and accuracy, but not for syntactic complexity. This may be explained in the following way: whether learners use simple or complex syntactic structures may simply not have an impact on the perception by raters that learners are being more or less communicatively adequate. On the contrary, the range of vocabulary employed by learners as well as the accuracy of the productions may be associated with the perception that they are also communicatively adequate. This is especially the case when proficiency is factored in, since results for accuracy show a moderately strong correlation with communicative adequacy. This finding also suggests that it may be worthwhile to take a closer look at the results of each individual learner.

A similar picture emerged with respect to the third research question, concerning the relationship between linguistic complexity as assessed by the raters on a six point Likert scale and linguistic complexity as assessed by general measures, that is, there were significant correlations for lexical diversity and accuracy but not for syntactic complexity. Results of structural complexity did not trigger any significant correlations and they also suggest that learners, in general, did not use highly complex structures. It is an issue whether more finegrained measures would capture instead any differences in structural complexity (like the elaboration of the noun phrase and the use of subordinate clauses, which were mentioned in the introduction). The results seem to imply that the decisions by raters to grade the students' general linguistic complexity may have been influenced more by the range of vocabulary they used and the accuracy of their productions than by the linguistic complexity of the text, as shown by the moderately strong correlations that were obtained for lexical diversity and accuracy. Accuracy in particular seems to determine the teachers ratings as the correlations tend to decrease when we take into account the proficiency level of the students.

The answers to the three research questions have broadened our view with respect to the nature of the relationship between communicative adequacy and linguistic complexity in L2 writing. The results have given us some insight into the role of communicative adequacy in relation to linguistic complexity and into the assessment of language proficiency by means of raters versus the use of general measures known from SLA research literature. We should, of course, take account of the limitations of our study. One such limitation is that although the participants were submitted to two tasks we only used one task type. But the problem that is troubling us most is the question of what makes raters decide whether a text is considered to be communicatively adequate or not. Interviews with raters and think aloud protocols while raters are judging texts might give us more insight into the motives raters are using to determine the communicative adequacy of a text.

Many other issues remain to be investigated. These include for instance the comparison between the written production of the L2 learners compared to that of the control group of native speakers. Another interesting comparison concerns the differences regarding the relationship between communicative adequacy and linguistic complexity in the three target languages: Dutch L2, Italian L2 and Spanish L2. It would perhaps also be worthwhile to consider carrying out in-depth analyses of specific features of L2 writing, such as the construction of the noun phrase and the verbal phrase or the use of subordinated clauses. But what may be by far the most tempting endeavour is to further explore the role of communicative adequacy in relation to complexity, accuracy and fluency. Another challenge is to investigate whether there are other, more 'objective' ways of measuring communicative adequacy. However, attempts to grasp this notion are still in their infancy.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659–663.

Babaii, E., & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System, 29*, 209–219.

Council of Europe. (2001). *Common European framework of references for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2007). The effect of task complexity on fluency and functional adequacy of speaking performance. In S. Van Daele, A. Housen, M. Pierrard, F. Kuiken, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 53–63). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (in press). The effect of task complexity on native and non-native speakers' functional adequacy, aspects of fluency, and lexical diversity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Investigating complerxity, accuracy and fluency in SLA.* Amsterdam: John Benjamins Publishing Company.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantatative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49.

Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System, 27*(1), 79–89.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14*(1), 47–84.

Little, D. (2007). The Common European Framework of reference for language perspectives on the making if supranational language education policy. *The Modern European Language Journal, 91*, 644–652.

McNamara, T. F., & Roever, C. (2007). *Testing: The social dimension.* Malden, MA/Oxford UK: Blackwell Publishing.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. Special issue, complexity, accuracy and fluency (CAF) in second language acquisition. *Applied Linguistics, 30*(4), 590–601.

Seliger, H., & Shohamy, E. (1989). *Second language research methods.* Oxford University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity.* Honolulu, Hawai'i: University of Hawai'i Press.

APPENDIX 1. Task

Every month your favourite newspaper invites its readers to have a say in what will be the leading article for the monthly supplement. This time the Editorial Board has come up with three suggestions: 1) global warming, 2) physical education 3) animal experiments.

Out of these three suggestions one has to be selected. The selection is made by a Readers' Committee. Every member of the committee has to write a report to the editors in which she/he states which article should be selected and why. On the basis of the arguments given by the committee members the Editorial Board will decide which article will be published on the front page. This month you have been invited to be a member of the Readers' Committee. Read the brief descriptions of the suggestions for articles below. Determine which article should be on the front page and why. Write a report in which you give at least three arguments for your choice. Try to be as clear as possible and include the following points in your report:

- which article should be selected;
- what the importance of the article is;
- which readers will be interested in the article;
- why the editorial board should place this article on the front page of the Special Magazine (give three arguments),

You have 35 minutes available to write your text and you need to write at least 150 words (about 15 lines). The use of a dictionary is not allowed.

Suggestions for articles:

1. *Global warming*: there is an ongoing political and public debate worldwide regarding what, if any, action should be taken to reduce global warming.
2. *Physical education*: the government is launching a campaign in order to prevent people from becoming obese and to encourage them to move more.
3. *Animal experiments*: it is estimated that 50 to 100 million animals worldwide are used annually and killed during or after experiments.

APPENDIX 2. Levels of communicative adequacy

0	1	2	3	4	5	6
The participant does not communicate any relevant information	The information which the participant communicates scarcely describes the chosen organization but does not describe its objectives or the people who will benefit from it. The arguments are extremely simple (i.e., single statements not supported by other arguments). They are also not connected to the organization's goals and beneficiaries. And/or: The text lacks coherence, and it is very confusing. > an unsuccessful contribution	The information which the participant communicates briefly describes the chosen organization, and scarcely describes any of its objectives, or the people who will benefit from it. The arguments are simple (i.e., single statements not supported by other arguments) and poorly connected to the organization's goals and beneficiaries. And/or: The text lacks coherence, and it is confusing. > a weak contribution	The information which the participant communicates briefly describes the chosen organization, its objectives, or the people who will benefit from it. The arguments are simple (i.e., single statements not supported by other arguments) but connected to the organization's goals and beneficiaries. And/or: The text is somewhat coherent but a little effort needs to be made. The text meets the minimum requirements. > a moderately successful contribution	The information which the participant communicates clearly describes the chosen organization, its objectives, or the people who will benefit from it. The arguments are complex (e.g. single or multiple statements supported by other arguments) and connected to the organization's goals and beneficiaries. And: The text is coherent. > a successful contribution	The participant contributes complete information about the chosen organization, its objectives, or the people who will benefit from it. The arguments are complex and elaborate (e.g. single or multiple statements supported by other arguments) and clearly connected to the organization's goals and beneficiaries. And: The text flows smoothly, it is coherent, and is convincing. > a very successful contribution	The participant contributes very complete and precise information about the organization, its objectives, or the people who will benefit from it. The arguments are well developed and well structured, and also highly convincing. This information is very well connected with the organization's goals and beneficiaries. And: The text flows smoothly. It is well constructed, coherent and convincing. > a highly successful contribution

APPENDIX 3. Levels of linguistic complexity

0	1	2	3	4	5	6
The participant has a very limited and basic range of simple expressions and vocabulary related to factual information. Word choice is often wrong. Sentences are extremely simple. The participant shows only a limited control of a few simple grammatical structures and sentence patterns. And/or The mistakes in the text make comprehension nearly impossible. > a very poor text in terms of vocabulary, grammar, and orthography.	The participant uses a limited range of highly frequent words and expressions. Some instances of wrong word choice. Sentences are quite simple. The participant uses some simple grammatical structures correctly, but still makes a considerable number of mistakes, including both grammatical (e.g. agreement, verb tenses, prepositions) and orthographic. And/or The mistakes in the text make comprehension difficult. > a rather poor text in terms of vocabulary, grammar, and orthography.	The participant uses a limited range of highly frequent words and expressions with a few instances of wrong word choice. Sentences are simple. The participant uses simple grammatical structures with some mistakes, including both grammatical (e.g. agreement, verb tenses, prepositions) and orthographic. And/or The many mistakes in the text require a little effort from the reader to comprehend the text. > a poor text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of highly frequent words and expressions but there are no instances of wrong word choice. Sentences are simple but they display some complex structures (e.g. relative clauses). The participant uses simple grammatical structures with the odd mistake (e.g. wrong preposition). And There are a few basic/ elementary mistakes that do not prevent comprehension. > an acceptable text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of both high and low frequency words and expressions and there are no instances of wrong choice and words and expressions are appropriate in general. Sentences are moderately complex (e.g. relative clauses, less frequent structures) with no mistakes. And There are minor mistakes or mistakes which are the consequence of trying to use more complex language. > a well written and accurate text in terms of vocabulary, grammar, and orthography.	The participant uses a wide range of low frequency and specific words and expressions and there are no instances of wrong word choice which is appropriate to the context. Sentences are quite complex (e.g. relative clauses, infrequent structures) with no mistakes. And There are almost no mistakes in the text and they may be the consequence of trying to use more complex language > a very well written and accurate text in terms of vocabulary, grammar, and orthography.	The participant uses complex (infrequent and rich) and specific vocabulary and expressions that are especially appropriate to the context. Sentences are highly complex, combining many different grammatical structures. And The text contains only the odd mistake. > an extremely well written and accurate text in terms of vocabulary, grammar and orthography.