TOWARDS CRITERION VALIDITY IN CLASSROOM LANGUAGE ANALYSIS: METHODOLOGICAL CONSTRAINTS OF METADISCOURSE AND INTER-RATER AGREEMENT[1]

Douglas Altamiro CONSOLO[2]

■ ABSTRACT: This paper reports on a process to validate a revised version of a system for coding classroom discourse in foreign language lessons, a context in which the dual role of language (as content and means of communication) and the speakers' specific pedagogical aims lead to a certain degree of ambiguity in language analysis. The language used by teachers and students has been extensively studied, and a framework of concepts concerning classroom discourse well-established. Models for coding classroom language need, however, to be revised when they are applied to specific research contexts. The application and revision of an initial framework can lead to the development of earlier models, and to the re-definition of previously established categories of analysis that have to be validated. The procedures followed to validate a coding system are related here as guidelines for conducting research under similar circumstances. The advantages of using instruments that incorporate two types of data, that is, quantitative measures and qualitative information from raters' metadiscourse, are discussed, and it is suggested that such procedure can contribute to the process of validation itself, towards attaining reliability of research results, as well as indicate some constraints of the adopted research methodology.

■ KEYWORDS: Discourse analysis; classroom discourse; metadiscourse; validation; reliability.

1 An earlier version of this paper was presented in the poster session at the XI AILA World Congress, Jyväskylä, Finland, 4-9 August 1996.

2 Departamento de Letras Modernas – Instituto de Biociências, Letras e Ciências Exatas – UNESP – 15054-000 – São José do Rio Preto – SP. E-mail: dconsolo@lem.ibilce.unesp.br.

Categorization of discourse acts in classroom interaction

This paper reports on a process to validate a system for coding classroom discourse in foreign language lessons. The categories of analysis for teacher talk and student speech proposed for validation in this study were adapted from a range of models and studies on classroom discourse,[3] mainly from the categories devised in earlier models to classify and describe different communicative aims and pedagogical purposes of teacher talk and student speech. The typology of discourse categories dealt with in this paper is based on the work of Consolo (1996), an extended and revised version of a model originally proposed by Sinclair & Coulthard (1975, 1992) for analysing classroom interaction.

The units of analysis at the lowest rank of discourse are called "communicative acts". Beyond their formal syntactic and phonological properties, acts are linguistic items at the level of discourse, that is, they are analysed according to the functional properties with what speakers use acts for.

One assumption in this categorization of discourse elements is that the discourse value of an act depends on what acts have preceded it, what are expected to follow, and the way acts relate to each other. Such sequential relationship is considered to analyse the discourse structure by means of categories defined not only structurally, but also for classroom-specific situations. The sociolinguistic context of formal EFL lessons, including non-verbal behaviour and paralinguistic aspects of the environment determine the structural and functional categories in classroom discourse (henceforth CD). The relevant "non-linguistic" factors to influence the production and analysis of classroom language – the social conventions of the environment and the shared experience of the speakers – are treated in the Sinclair & Coulthard (henceforth S&C) model (1975) as "situation". Such factors imply an analysis of language and interaction in EFL lessons based on the awareness of social, pedagogical, psychological and linguistic aspects of classroom interaction.

Acts constitute the next higher rank of discourse categories, defined as "moves". These categorise the typical interactional Initiation-Response-Follow up (IRF) structure of lessons, which predicts student moves as responses (R) to the teacher's initiations (I). Student speech is usually evaluated by the teacher, in follow-up (F) moves. A differentia-

3 For a detailed report on the development of the coding system dealt with here, see Consolo, 1996, p.144-87 (Chapter 4)

tion is made in the system proposed by Consolo (1996) between TT and student speech at the rank of moves by using the labels Is, Fs and Rt for initiations (Is) and follow-up moves (Fs) produced by students, and responses given by the teacher (Rt), in order to analyse the specific contributions of teachers and students to CD.

The interactional unit is defined as an exchange, which is determined by the occurrence of an I or Is move. Exchanges combine to form transactions. The identification of a transaction boundary usually coincides with topic change, that may be marked by boundary moves such as Frame (Fr) and Focus (Fo). Transactions differ from exchanges in that while it is possible to establish internal patterns and a categorization for exchanges,[4] little can be said ab|out the patterns for the internal structure of transactions. Transactions combine to form lessons, the highest unit on the rank scale for analysing CD.

The acts which categorise the functions of classroom language, as produced by teachers and students, are presented below:[5]

1. Maker {mrk}
2. Starter {str}
3. Elicitation {eli}
4. Comprehension check {cp-chk}
5. Confirmation check {cf-chk}
6. Clarification {cla}
7. Directive {dir}
8. Informative {inf}
9. Comment {com}
10. Clue {clu}
11. Model {mdl}
12. Bid {bid}
13. Nomination {nom}
14. Acnowledge {ack}
15.1. Affirmative reply {y-rpl}
15.2. Negative reply {n-rpl}
15.3. Choice reply {c-rpl}
15.4 Repetition reply {rp-rpl}
15.5 Informative reply {i-rpl}
15.6. Offer reply {o-rpl}
16. Rect {rea}
17. Protest {pro}
18. Correction {cor}
19. Evaluate {evl}
20. Metastatement {mst}
21. Apology {apl}
22. Tank {thk}
23. Encouragement {ecg}
24. Conclusion {con}
25. Terminate {ter}
26. Greeting {grt}
27. Parting {prt}
28. Aside {asi}
29. Translation {trl}
30. Rhetorical aquestion {rtq}

4 Various patterns, such as IR and I{RF}^{n}, derived from the basic IRF structure are presented in Sinclair & Couthhard (1975) and Consolo (1996).

5 For the definitions of the acts, see Consolo, 1996, p.170-85.

The next section reports on a process to validate the revised version of the S&C model proposed by Consolo (1996), focusing at the rank of the communicative acts listed above. The validation of these revised categories of analysis is essential to support the results and further implications of such a study, since its achievements will have been reached through criteria other than those prescribed by the original model.

Validation of categories of analysis

The validity and reliability of observational procedures, and of categories of analysis, are key issues in research. A major aim in classroom research is, according to Chaudron (1988, p.23),

> to produce descriptions and interpretations of classroom events, and the relationships between them, that will be identified by others as real and meaning ful for teachers, learners, and learning.

The assumption above is followed in this study with regard to the description of how teachers and students engage in verbal interaction in EFL lessons, and to the extent which classroom observation, recorded lessons and transcripts can reveal the meanings and goals of that type of interaction (Stubbs, 1983).

Validity and reliability in data analysis are focused on and investigated here for a number of reasons. Firstly, since both qualitative and quantitative approaches were followed to collect and analyse the data corpus assembled by Consolo (1996), the research design and methodology included different procedures and research instruments. As a result, different types of information, for example, classroom data obtained by means of audio-recordings and by direct observation, were combined in lesson transcripts. The way such information was gathered and put together by the researcher alone is likely to favour deviations between the nature of phenomena, as they occur in the field, and their interpreted versions (Lampert & Ervin-Tripp, 1993; Ochs, 1979). Recordings reveal classroom phenomena with a considerably high degree of reliability. Data from field notes and transcripts, on the other hand, are affected by intervening factors concerning the observer's interpretation of phenomena, as well as his or her interpretation of recordings while lesson transcripts are produced. These factors can make research outcomes subjective and less reliable, thus limiting generalization of results (Lampert & Ervin-Tripp, 1993; Tesch, 1990).

Two "validation sessions", referred to as validation sessions 1 (VS1) and 2 (VS2) were carried out so as to verify the validity and the reliability of part of the system of categories used to code lesson transcripts. Due to practical constraints in having all categories of data analysis validated, the decision was to proceed with using part of the coding system in the same way those categories had been used to code data in earlier studies. Categories such as exchanges and moves[6] have therefore not been fully validated neither in VS1 nor in VS2. Raters were asked to focus on those categories that had been subject to adjustment and redefinition, as from the pilot study and the preliminary data analysis carried out by Consolo (1996),[7] all at the rank of acts.

VS1 and VS2 aimed at (1) the training of external raters to use the system on lesson transcripts, and (2) to have the raters code samples of data, that is, extracts from lesson transcripts, by following specific descriptions of the categories of analysis.

For VS1, eight professionals (teachers and/or research students) working in the areas of ELT/ELT were contacted regarding their voluntary contribution to validate the system of categories of analysis in this investigation. However, only four of them were actually present in the session. The "external raters" in VS1 are referred to as rater 1, rater 2, rater 3 and rater 4.

Rater 1 and rater 2 are native speakers of English, and had been working in ELT for many years when they took part in VS1. Rater 1 holds an MA in TEFL, and had been formally exposed to some of the theoretical background reviewed for this study while she did her MA. Rater 2 was working towards his MA in TEFL when the validation session was carried out. He had previously worked in ELT in Brazil, which means not only does he speak some Portuguese, but also he is familiarized with the teaching contexts investigated here.

Rater 3 and rater 4 were working on their doctoral theses at the time. Rater 3 speaks English as L2, and his research was in the area of spoken English. Rater 4 is a non-native speaker of English. Both of them are proficient in the English language.

Both the training of raters and the coding of data samples were carried out on the same occasion due to the constraints imposed on counting on voluntary external cooperation. A common problem is to associate the

6 As for the distinction between moves in TT (I, Rt and F) and in student speech (Is, R and Fs), the characteristics of those moves were clearly understood by raters in both VS1 and VS2.

7 See Consolo, 1996, p.105-12.

practicalities of having the required phases of training raters and validating categories of analysis on different occasions, with the little availability offered by most professionals who match the necessary standards to be external raters.[8]

If "well-trained and like-minded coders" (Lampert & Ervin-Tripp, 1993, p.196) are desirable, favourable conditions are to be established so as to train raters properly. The training of raters 1 - 4 in VS1 was concentrated in a short time slot; that one or possibly two raters did have not learned enough about the categories of analysis is strongly suggested by unsatisfactory results in their using the codes which resulted in low levels of agreement, as reported below.

The preparation of materials

The samples from lesson transcripts used to validate the categories of analysis included both plain and coded transcripts; these had been randomly selected from the corpus, according to the criteria explained below.

Samples were from ten whole lessons previously transcribed and coded by the researcher. The turn was initially adopted as an interactional reference within the transcribed lessons, since it can be clearly identified as a unit of spoken language and verbal interaction (this resembles the path followed by Sinclair and Coulthard, 1975, towards the definition of moves), to be coded in terms of moves and acts. More specifically, ten sequences of twenty turns each were selected from the total of 6558 turns contained in ten lesson transcripts. The total number of turns was adjusted to 6560, from which 328 twenty-turn sequences were obtained. For example, turns 001-020 were from "Lesson 1" and stand as "sequence 001". Turns 6540 - 6558, from "Lesson 10",[9] stand as "sequence 328".

The ten sequences randomly chosen (by referring to a table of random numbers) are indicated in Table 1 below:

8 Those factors apply to limited availability under the conditions of "voluntary participation" in research. The problem may be eliminated by having participants' availability and professional commitment in return to paid work.

9 The labels "Lesson 1 - Lesson 10" for samples used in the validation sessions are distinct from the labels LES 1, LES2, LES3 and so on, referring to lessons taught by each teacher.

Table 1 – Validation sessions: sample sequences (from audio-recorded lessons)

Teacher[10]/Lesson	Turns	Sequences (numbering 001-328)
NS1/"Lesson 1"	400 - 420	1 (020)
	421 - 440	2 (021)
	441 - 460	3 (022)
NNS1/"Lesson 2"	021 - 040	4 (028)
	281 - 300	5 (041)
NNS4/"Lesson 8"	141 - 160	6 (229)
	581 - 600	7 (251)
	721 - 740	8 (258)
NS2-ADV3/"Lesson 9"	001 - 020	9 (268)
	601 - 620	10 (298)

Further decisions concerning the use of the selected samples such as which samples would be more suitable for training raters or for validating the categories, and the actual operationalization of the training and the coding phases in the validation session, led to the final choice of six samples, three of which are presented in Appendices 1-3.

Working materials (booklets with definitions[11] and examples, handouts with the selected sequences) were prepared and used for training raters, and for the checking on the reliability of the categories of analysis.

The training phase

A fully coded sample (Appendix 1) was the departure point for presenting the codes for transcription and examples of the categories of analysis (exchanges, moves and acts). The sample in Appendix 2 was used for training raters to do the coding of moves and acts. The audio tapes were also played in the validation sessions so as to illustrate varia-

10 Teachers involved in Consolo's (1996) study are native speakers (NS) and non-native speakers (NNS) of English.

11 As presented in Consolo, 1996, p.171-86. In VS1, raters had to get accquainted with the definitions during the first part of the session, and it is arguable whether they were able to grasp the definitions of the acts in order to apply such categories in the coding of data samples. The second group of raters (VS2) had the opportunity to study the definitions at their will between the first and the second day.

tions in meaning conveyed by intonation contours[12] and provide raters with a more reliable account of the phenomena under investigation.

Results: Validation Session 1

The subcategories of acts dealt with in VS1 and levels of agreement between each two raters who have rated the occurrences of the same acts are presented in Table 2:

Table 2 – Agreement between raters in VS1

Range of acts	Raters	Level of agreement (p_o)	Coefficient of agreement: κ
SUBGROUP I	rater 1 - rater 2	34.6% (p_o= 0.34)	0.25
{mrk}, {str}, {eli}, {cp-chk}, {cf-chk}, {cla}, {dir}, {inf}, {com}, {clu}, {mdl}	researcher - rater 1	74% (p_o= 0.74)	0.65
	researcher - rater2	38.4% (p_o= 0.38)	0.29
SUBGROUP II	rater 3 - rater 4	51.4% (p_o= 0.51)	0.46
{bid}, {nom}, {ack}, {y-rpl}, {n-rpl}, {c-rpl}, {rp-rpl}, {i-rpl}, {rea}, {pro}	researcher - rater 3	71.4% (p_o= 0.71)	0.68
	researcher - rater 4	64% (p_o= 0.64)	0.58

The procedure for validating the acts was to have two raters categorizing the same data samples independently, and then determine the degree and significance of the raters' agreement (Cohen, 1969; Lampert & Ervin-Tripp, 1993). Following Cohen (1969), the proportion of cases about which the raters agreed (nominal scale agreement) is determined by the calculation of p_o. For example, the inter-rater agreement between the researcher and rater 1 for acts ranging between "markers" and "models", as shown in Table 3 below, is given by

$$p_o = 0.27 + 0.06 + 0.04 + 0.02 + 0.29 + 0.02 + 0.04 = 0.74$$

A certain amount of agreement by chance is, however, expected. This can be determined by multiplying the probabilities of the margi-

12 Intonation has not been indicated in transcripts, except for "?" for interrogatives, and standard intonation patterns such as [RISE], [FALL], [FALL-RISE] and [RISE-FALL].

nals. For acts 1 (marker) - 11 (model), the researcher placed 0.08 of the cases under the category "confirmation checks" (cf-chk), while rater 1 placed 0.04 of the cases in this category. The expected chance agreement for cf-chk is then (0.08)(0.04) = 0.0032. Values for chance agreement are the parenthetical entries along the agreement diagonal in Table 3. The proportion of agreement to be expected by chance, p_c, is found by adding the parenthetical values:

$$p_c = 0.1089 + 0.0084 + 0.0032 + 0.0008 + 0.1369 + 0.0008 + 0.0024 = 0.2614$$

The coefficient of agreement, κ, is the proportion of agreement after agreement by chance is removed from consideration, which can be obtained as follows:

$$\kappa = \frac{p_o - p_c}{1 - p_c} = 0.65$$

Values of κ ranged 0.25 - 0.65 for acts in subgroup I and 0.46 - 0.68 for subgroup II. Given the minimal amount of training given to raters 1 - 4 and considering the observable characteristics of their handling of categories and data samples during VS1, the levels of agreement for acts dealt with raters 3 and 4 were slightly higher than the ones for subgroup I (and especially for rater 2) maybe due to those raters' previous knowledge of the S&C model and current involvement in research. Rater 2's low levels of agreement, the least experienced member in the group as far as both discourse categories and research procedures are concerned, corroborate this conclusion.

Table 3 – Inter-rater agreement Researcher-Rater1(VS1) – Across: Researcher / Down: Rater1 Acts 01 (marker) – 11 (model) – Total: 52 occurrences

Selected acts	1. mrk	2. str	3. eli	4. cp-chk	5. cf-chk	6. cla	7. dir	8. inf	9. com	10. clu	11. mdl	p_i rater 7
1. mrk												
2. str												
3. eli			0.27 (0.1089)					0.06				0.33
4. cp-chk			0.02	0.06 (0.0084)	0.04	0.02						0.14
5. cf-chk					0.04 (0.0032)							0.04
6. cla												
7. dir		0.02					0.02 (0.0008)					0.04
8. inf		0.02	0.04					0.29 (0.1369)		0.02		0.37
9. com									0.02 (0.0008)		0.02	0.04
10. clu										0.04 (0.0024)		0.04
11. mdl												
N/A								0.02				0.02
p_i researcher			0.33	0.06	0.08	0.02	0.02	0.37	0.02	0.06	0.02	$\sum p_i$ = 1.00

$p_o = 0.74$

$p_c = 0.2614$

$\kappa = 0.65$

Validation Session 2

Because the levels of agreement achieved in VS1 were not considered satisfactory, and given the chances that such results may have been negatively affected by the conditions under which VS1 was conducted, a second session was carried out. The raters in VS2 were all MA TEFL students, with the exception of one member of the academic staff who volunteered to join the group. The six raters in VS2 are referred to as rater 5 - rater 10. Unlike raters 1 - 4, raters in VS2 (except for the staff member) had not been explicitly exposed to the theoretical background for categories of discourse analysis. Thus raters in VS2 needed sound teaching and practice in using the codes before attempting to rate the set of acts.

VS2 was actually conducted on three different days so as to allow for appropriate training of raters and satisfactory rating of the coding system. The first two meetings (days one and two) were entirely dedicated to the presentation of the study and a description of the coding system, by the researcher, to the group of raters. The same samples from lesson transcripts used to train raters in VS1 were coded and discussed, and raters 5 - 10 had more time than raters 1 - 4 to understand the categories and procedural aspects of coding followed by the researcher.

On the third day, the actual rating of the categories at the rank of acts was carried out. Raters were randomically grouped in pairs, and each pair dealt with a previously determined subgroup of acts, as in VS1. In VS2 nearly all the acts were rated,[13] as opposed to VS1, in which the number of raters was too small to dealt with all types of acts.The procedure of rating subgroups of acts was adopted so as to facilitate the handling of definitions to be followed by raters. In this way, categories for occurrences of specific acts that may lead to double-labelling, for example comments and informatives, were simultaneously tested by two raters working on the same samples. Appendix 3 illustrates how the same sample was prepared to be handled by different raters, each group dealing with a predetermined set of acts. Gaps were prepared for the labelling of acts within the range being dealt with each pair of raters.

13 Four types of acts – greetings, partings, asides and translations – have not been validated in this study due to the reduced number of raters and the short amount of time to deal with the large number of categories. It has been assumed that, since those categories are drawn from similar studies on classroom behaviour and their discourse functions are less prone to ambiguous interpretation, they would be the ones to be left out.

Because this procedure resembles the technique of testing language by focusing attention on one point (grammar or vocabulary, for example) at a time, the coding of acts was operated under the principles of a "discrete point test" (Oller Jr., 1979), and in which raters chose the categories from a pre-established set of "alternatives", as in a multiple choice test.

The levels of agreement for each set of acts in VS2 are presented in Table 4 below:

Table 4 – Agreement between raters in VS2

Range of acts	Raters	Level of agreement	Coefficient of agreement: κ
SUBGROUP I	rater 5 - rater 6	44.2% (p_o= 0.44)	0.41
{mrk}, {str}, {eli}, {cp-chk},			
{cf-chk}, {cla}, {dir}, {inf},	researcher - rater 5	37% (p_o= 0.37)	0.25
{com}, {clu}, {mdl}	researcher - rater 6	54% (p_o= 0.54)	0.41
SUBGROUP II	rater 7 - rater 8	80% (p_o= 0.8)	0.75
{bid}, {nom}, {ack}, {y-rpl},			
{n-rpl}, {c-rpl}, {rp-rpl},	researcher - rater 7	97% (p_o= 0.97)	0.96
{i-rpl}, {rea}, {pro}	researcher - rater 8	83% (p_o= 0.83)	0.79
SUBGROUP III	rater 9 - rater 10	40% (p_o= 0.4)	0.31
{ack}, {cor}, {evl}, {mst}, {apl},			
{thk}, {ecg}, {con}, {ter}, {grt}	researcher - rater 9	65% (p_o= 0.65)	0.51
{prt}	researcher - rater 10	45% (p_o= 0.45)	0.24

The highest levels of agreement for all the rated acts, in both VS1 and VS2, were therefore 74% for acts in subgroup I (κ=.65), as shown in Tables 2 and 3, 97% for acts in subgroup II (κ=.96) and 65% for acts in subgroup III (κ=.51), as shown in Table 4.

Levels of agreement reached in VS1 and VS2 are, on the whole, not as high as the levels reached in earlier studies in which categories for analysing lesson transcripts were validated, as for example in Mitchell et al. (1981) and Ramirez et al. (1986).[14] The results reached here may have had the intervention of the following factors. Firstly, the complexity of the system of discourse exchanges and acts, with its large number of proposed categories, especially at the level of acts. The specifics of each definition that, in some cases, distinguish acts that have similar communicative functions, have to be carefully grasped by coders. This seems

14 The procedure for determining reliability on the coding of lesson transcripts in the study conducted by Ramirez et al. (1986) was developed by C. Chaudron and others at the University of Hawaii, as reported in Chaudron (1988, p.24).

to be feasible, however, only after extensive theoretical and methodological familiarity with the system and the coding procedure. Secondly, lack of familiarity with the system probably led most raters to misinterpret definitions and/or have difficulty in coding acts that are expected to be problematic due to their similar functions. These are, for example, clarifications (cla), confirmation checks (cf-chk) and comprehension checks (cp-chk); clues (clu), informatives (inf) and starters (str); and acknowledges (ack) and evaluates (evl). Cases in which the researcher and raters disagreed on those acts are shown in Table 5 below:

Table 5 – Acts: raters' disagreement in VS1 and VS2

Subgroup I		
Case n.	Researcher	Rater 1
(12)	cf-chk	cp-chk
(15)	str	inf
(20)	str	dir
(21)	eli	inf
(23)	eli	inf
(24)	inf	eli
(27)	inf	eli
(28)	inf	eli
(31)	inf	–
(37)	com	cor
(38)	mdl	cor
(43)	eli	cp-chk
(44)	clu	inf
(45)	cla	cp-chk

Subgroup II		
Case n.	Researcher	Rater 7
(21)	pro	inf

Subgroup III		
Case n.	Researcher	Rater 9
(4)	ack	i-rpl
(6)	ter	ack
(13)	ack	evl
(15)	ack	evl
(17)	ack	HES
(18)	ack	evl
(20)	ack	evl

A third factor affecting the levels of agreement was misinterpretation, by raters, of definitions and coding instructions, leading to discrepancies in cases such as (21), (23), (24), (27) and (28) for subgroup I, in which the researcher and rater 1 disagreed on elicitions (eli) and informatives (inf); and in cases (4) and (17) for subgroup III, in which rater 9 did not follow the instructions and used categories other than the actas in subgroup III. These drawbacks can be linked to the conditions under which VS1 was carried out, for despite theexperience of raters 1-4, as reported in their profiles above, they were not fiben enough training to deal with the categories proposed here.

Nevertheless, high levels of agreement are possible, as shown by the results reached by rater 7 (subgroup II) in comparison to the researche's (κ=.96), as shown in Table 6 below:

Table 6 – Inter-rater agreement Researcher-Rater 7 (VS2) – Across: Research/Down: Rater 7 Acts 12 (bid) – 17 (protest) – Total: 30 occurrences

Selectd acts	12. bid	13. nom	14. ack	15.1. y-rpl	15.2. n-rpl	15.3. c-rpl	15.4. rp-rpl	15.5. i-rpl	16. rea	17.pro	p_i rater 7
12. bid											
13. nom		0.03 (0.0009)									0.03
14. ack			0.24 (0.0576)								0.24
15.1. y-rpl				0.14 (0.0196)							0.14
15.2. n-rpl					0.03 (0.0009)						0.03
15.3. c-rpl						0.24 (0.0576)					0.24
15.4. rp-rpl							0.03 (0.0009)				0.03
15.5. i-rpl								0.26 (0.0676)			0.26
16. rea											
17.pro											
N/A										0.03	0.03
p_i researcer		0.03	0.24	0.14	0.03	0.24	0.03	0.26		0.03	$\Sigma\, p_i$ = 1.00

p_o = 0.97

π_c = 0.2051

κ = 0.96

Rater 7 was the one collaborator who probably best learned how to apply the categories by studying them carefully between the meetings.[15] This corroborates the conclusions reached above as for the insufficient preparation of the other raters.

Conclusion

The process to validate a system for coding classroom discourse in FL lessons reported in this paper has proved useful for a better understanding of validity and reliability of a given set of discourse categories. The categories were applied in a context in which the dual role of language and the speakers' specific pedagogical aims lead to a certain degree of ambiguity in language analysis, and that is one reason why the level of inter-rater agreement was not particularly high.

The validation of communicative acts confirmed the complexity of the coding system, and the ambiguities found in the definitions of some acts. Such ambiguity is also a consequence of the ambiguity in classroom communication, as pointed out above. For example, the tendency of usually interpreting teachers' acts in F moves as evaluates blurs their other functions, such as acknowledging student speech, providing information or terminating an exchange.

The overall set of procedures followed to validate the coding system presented here (selection of data samples, training of raters, data coding and results), despite the constraints faced in this study, may be seen as guidelines for conducting research under similar circumstances, and attempting to obtain higher levels of reliability in classroom language analysis.

CONSOLO, D. A. Em busca da validade criterial na análise da linguagem de sala de aula: limitações metodológicas do metadiscurso e da concordância entre avaliadores. *Alfa (São Paulo)*, v.43, p.113-134, 1999.

▪ *RESUMO: Este artigo relata um processo para validar um sistema de categorização do discurso em aulas de língua estrangeira, um contexto no qual o duplo papel da linguagem (conteúdo e meio de comunicação) e os objetivos*

15 According to rater 7's own statement.

pedagógicos específicos dos falantes geram um grau de ambigüidade nos significados dessa linguagem. A linguagem de sala de aula foi investigada, objetivando estabelecer um arcabouço de pressupostos e categorias de análise do discurso pedagógico. As categorias propostas no estudo tratado (Consolo, 1996, p.144-87) foram adaptadas a partir de uma variedade de modelos para a análise do discurso e da interação em sala de aula. Modelos propostos para tal análise necessitam, entretanto, de revisões e adaptações ao serem utilizados em contextos diversos daqueles para os quais foram utilizados originalmente e, nesse processo de (re-)adaptação, redefinem-se categorias existentes, bem como criam-se novas categorias. Nesse processo, faz-se necessário validar um modelo recriado, para garantir a cientificidade do estudo realizado e suas implicações. Apresentamos os procedimentos utilizados para validar o modelo de análise utilizado por Consolo (ibidem), incluindo-se a seleção dos dados, o treinamento de pesquisadores-participantes, a codificação das categorias e os resultados do estudo. Sugere-se que tais procedimentos contribuam para a validade e confiabilidade de resultados de pesquisas, enquanto revelam limitações decorrentes da metodologia adotada.

- *PALAVRAS-CHAVE: Análise do discurso; discurso de sala de aula; metadiscurso; validação; confiabilidade.*

References

CHAUDRON, C. *Second language classrooms*. Cambridge: CUP, 1988.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v.1, n.20, p.37-46, 1969.

CONSOLO, D. A. *Classroom discourse in language teaching*: a study of oral interaction in EFL lessons in Brazil. Cals, 1996. Tese (Doutorado) – The University of Reading.

LAMPERT, M. D., ERVIN-TRIPP, S. M. Structured coding for the study of language and social interaction. In: EDWARDS, J. A., LAMPERT, M. D. *Talking data*: transcription and coding in discourse research. Hillsdale, NJ: Lawrence Erlbaum, 1993. p.169-206.

MITCHELL, R., PARKINSON, B., JOHNSTONE, R. *The foreign language classroom*: an observational study. Stirling: University of Stirling, 1981. (Stirling monographs, 9)

OCHS, E. Transcription as theory. In: OCHS, E., SCHIEFFELIN, B. B *Developmental pragmatics*. New York: Academic Press, 1979. p.43-72.

OLLER Jr, J. W. Discrete Point, Integrative, or Pragmatic Tests. In: *Language Tests at School*: A Pragmatic Approach. London: Longman, 1979. p.36-73.

RAMIREZ, J. D. et al. *First year report*: longitudinal study of immersion programs for language minority children. Arlington: SRA Technologies, 1986.

SINCLAIR, J., COULTHARD, M. *Towards an analysis of discourse*: the english used by teachers and pupils. London: OUP, 1975.

_____. Towards an analysis of discourse. In: COULTHARD, M. *Advances in spoken discourse analysis*. London: Routledge, 1992. p.1-34.

SOARS, J., SOARS, L. *Headway advanced* (Student's book). Oxford: OUP, 1989.

STUBBS, M. *Language, Schools and classrooms*. 2.ed. London: Methuen, 1983.

TESCH, R. *Qualitative research*: analysis types and software tools. London: Falmer Press, 1990.

Appendix 1

CODED LESSON EXTRACT (NS1 – Lesson 2) – SAMPLE 1 FOR GROUP TRAINING – Turns: 396 - 424[16]

Information about the lesson: T and STS are discussing the students' impressions about the characters in the first part of a short story ("The Hitch-hiker", by Roald Dahl). The STS had already listened to the extract in class and followed the text in their coursebooks. Since only the beginning of the story is known by the STS, the T is working on predictions and suggestions for the development of the narrative. Setting: whole-class interaction.

(turns - speakers) (utterances)			(acts)	(moves)
396 - St3	/you/	{FS}		ls1
	/+ you're (speaking) abaut three (possible)/			
397 - T:	/oh + ok +/	{mrk}		Rt1(ls1)
	/so + one possibilitiy is + this man is a thief/	{i-rpl}		
	this man is a thief/			[151]
397 - T:	/ + so + what happens next?/	{eli}		I1
398 - St:	/ + (UNINT)(a gun)/	{i-rpl}		R1(I1)
399 - T	/he/	{FS}		I1 [152]
	/ + he steals the car:?/	{cla}		
	/ + he pulls out a gun [RISE]/		{cf-chk}	
	the says + drive me + to London [RISE]			
	/ + or he (tells) the guy to get out of the car + and	{clu}		
	drives off with his car + uhm?/		{eli}	
400 - St:	(UNINT)	N/A		(R1(I1))
401 - T	/ok +/	{mrk}		F1(R1)
	/that a possiblity/		{com}	
401 - T:	/ + anything else?/	{eli}		[BE1/152] I1
402 - St4:	/(maybe) + they could start talking abaut the car + er:	{i-rpl}		R1(I1)
403 - T:	who/	{FS}		I1 [153]
	/ + who would {star [RISE] }/	{eli}		
404 - St4:	/[the hitch-hiker]/	{i-rpl}		R1(I1)
	/and er + he (tries to)/	{i-rpl}		
405 - T:	/oh +/	{ack}		F1(R1)
	/(UNINT) + you have a:=/	{inf/407}		
406 - St:	(UNINT)	N/A		(R2(I1))
407 - T:	/=amazing car/	{405/inf}		F1(R1)
408 - St4:	/yeah/	{ack}		Fs1(F1)
	/+ and (UNINT)/	N/A		[BE1/152]

16 Only turns 396 – 408 have been reproduced in this appendix due to reasons of space.

Appendix 2

CODED LESSON EXTRACT (NNS4 – Lesson 1) – **SAMPLE 2 FOR GROUP TRAINING** – Turns 563 - 598[17]

Information about the lesson: T and STS are discussing a task done previously in pairs (a matching activity, "Children learn what they live", as in Headway Advanced).

CODING FOR MOVES AND ACTS

(turns - speakers) (utterances)	(acts)	(moves)
563 - T: /everybody agrees?/		
564 - St: /yes/		
565 - STS: /no/		
		[223]
566 - T: /{{CHUCK}} what did you have?		
567 - St: /I think + (UNINT) "to appreciate"/		
		[224]
568 - T: /with hostility?/		
569 - St: /yes/		
570 - St (UNINT)		
571 - St: /(UNINT)(fight)/		
572 - T: / [5] to fight?/		
/ + (that's possiblity)/		
		[225]
573 - St: /do you tink + that [RISE]/N		
{INTRRRUPTED TURN}		[226]
574 - St1: /to fight (could be) (UNINT)/		
575 - T: /yes/		
		[227]
576 - St1: /(to fight) + to: + to: [1] to (get) something/		
577 - St: /to get something that you want/		
578 - St: /yes/		
		[228]
579 - T: /but the author + shares the same opinion + that the child learns er + lives with hostility + he learns to fight/		
/ [3] I don't know/		
/ + he may be wrong/		
/ + I'm not saying that he is right [RISE-FALL]/		
/ [1] I'm just saying that + this is [(UNINT)]/		
580 - st: [(UNINT)		

17 Only turns 563 – 580 have been reproduced in this appendix due to reasons of space.

Appendix 3

CODED LESSON SAMPLE 2 (NNS1 – Lesson 1) FOR VALIDATING ACTS – Turns 007 - 031[18]

Information about the lesson: In the beginning of this lesson the T had set an exercise to be done in class. The task was to match sentences that contained prepositions to a list of cases that explained the uses of the prepositions in the sentences. At this point, the T is halfway through the correction of the exercise.

(Examples of gaps for acts validated by raters 1, 2, 5 and 6)

007 - T:	/[1] stephen/	{nom }	I1	
	/ + d/	1)_____		
008-ST1(*Stephen)	/er:/	{HES}	R1(I1)	
009 - T:	/"as it's one in the morning by then"/	2)_____	I1	[009] –
010 - St1:	/er:/	{HES}	R1(I1)	
	/ + four/	{i-rpl}		
011 - T	/ [2] four [RISE]/	3)_____	F1(R1)	
011 - T:	/ [3] do you agree?/	4)_____	I1	[BE1/009]
012 - St:	/no/	{n-rpl}	R1(I1)	
013 - T:	/no/	{ack}	F1(R1)	
013 - T:	/ [1] why four (then)?	5)_____	I1	[010] -
014 - St1:	/(UNINT)	N/A/016	(R1(I1))	
015 - T:	/wait wait/	6)_____	F1(R1)	
016-St1:	= (UNINT)/	014/N/A	(R1(I1))	
017 - T:	/(UNINT)/		N/A	[011] -
	/ + position?/	7)_____	I1	
	/ + place?/	8)_____		
	/ + position?/	9)_____		
018 - St1	/ [1] (er:)/	{HES}		
019 - T:	/or does it answer the question + when?	10)_____	I1	[012] –
020 - St:	/(UNINT)/	N/A	(R1(I1))	
021 - T:	/(UNINT)/		N/A	[BE1/012]
	/ + then + so [1] which is (the answer)?/	11)_____	I1	
022 - St1:	/ [1] (three)/	{i-rpl}	R1(I1)/Is1	
	/(right?)/	12)_____		[BE2/012]

18 A shorter sample is shown here (and therefore fewer gaps than those for the acts coded in this lesson sample) due to lack of space.

CODED LESSON SAMPLE 2 (NNS1– Lesson 1) FOR VALIDATING ACTS – Turns 007 - 031

(Examples of gaps for acts validated by raters 3, 4, 7 and 8)

007 - T:	/ [1] stephen/ / + d/	1)_____ {eli}	I1
008 - St1(Stephen):	/er:/	{HES}	R1(I1) [009]
009 - T:	/"as it's one in the morning by then"	{eli} {HES}	I1 R1(I1)
010 - St1:	/er:/ / + four/	2)_____	
011 - T:	/ [2] four {RISE}/	{cf-chk}	F1(R1) [BE1/009]
011 - T	/ [3] do you agree?/	{eli}	I1
012 - St:	/no/	3)_____	R1(I1)
013 - T:	/no/	4)_____	F1(R1) [010]
013 - T:	/ [1] why four (then)?/	{eli}	I1
014 - ST1:	/(UNINT) =	N/A/016	(R1(I1))
015 - T:	/wait wait/	{dir}	F1(R1)
016 - St1:	= (UNINT)/	014/N/A	(R1(I1)) [011]
017 - T:	/(UNINT)/ / + position?/ / + place?/ / = position?/	 {eli} {eli} {eli}	N/A I1
018 - St1:	/ [1] (er:)/	{HES}	[012]
019 - T:	/or does it answer the question when?	{eli}	I1
020 - St:	/(UNINT)/	N/A	(R1(I1)) [BE1/012]
021 - T:	/(UNINT)/ / + then + so [1] which is (the answer)?/	N/A {eli}	I1
022 - St1:	/ [1] (three) /(right?)/	5)_____ {cf-chk}	R1(I1)/Is1 [BE2/012]
023 - T:	/which one?/ / + one?/ (UNINT)	 {eli} {cf-chk}	I1 N/A [BE3/012]
023 - T:	/ + "she answers the door" + "looking a bit angry" +. "as it's one in the morning by then"/ / + by then + means what?/ / + by that + time/	{str} {eli} {inf}	I1 Rt1(I1) [BE4/012]

CODED LESSON SAMPLE 2 (NNS1 – Lesson 1) FOR VALIDATING ACTS – Turns 007 - 031

(Examples of gaps for acts validated by raters 9 and 10)

007 - T:	/ [1] stephen/	1)_____	I1
	/ + d/	{eli}	
008 - St1(Stephen):	/er:/	{HES}	R1(I1)
009 - T:	/"as it's one in the morning by then"	{eli}	I1 [009]-
010 - St1:	/er:/	{HES}	R1(I1)
	/ + four/	{i-rpl}	
011 - T:	/ [2] four [RISE]/	{cf-chk}	F1(R1)
011 - T	/ [3] do you agree?/	{eli}	I1 [BE1/009
012 - St:	/no/	{n-rpl}	R1(I1)
013 - T:	/no/	4)_____	F1(R1)
013 - T:	/ [1] why four (then)?/	{eli}	I1 [010]-
014 - ST1:	/(UNINT) =	N/A/016	(R1(I1))
015 - T:	/wait wait/	{dir}	F1(R1)
016 - St1:	= (UNINT)/	014/N/A	(R1(I1))
017 - T:	/(UNINT)/		N/A [011]-
	/ + position?/	{eli}	I1
	/ + place?/	{eli}	
	/ + position?/	{eli}	
018 - St1:	/ [1] (er:)/	{HES}	
019 - T:	/or does it answer the question when?/	{eli}	I1 [012]-
020 - St:	/(UNINT)/	N/A	(R1(I1))
021 - T:	/(UNINT)/	N/A	[BE1/012]
	/ + then + so [1] which is (the answer)?/	{eli}	I1
022 - St1:	/ [1] (three)	{i=rpl}	R1(I1)/Is1
	/(right?)/	{cf-chk}	
023 - T:	/which one?/		I1 [BE2/012]
	/ + one?/	{eli}	
	(UNINT)	{cf-chk}	N/A
023 - T:	/ + "she answers the door" + "looking a bit angry"	{str}	I1 [BE3/012]
	"as it's one in the morning by then"/		
	/ + by then + means what?/	{eli}	
	/ + by that + time/	{inf}	Rt1(I1)
023 - T:	/ [3] so is it one two three or four here?/ /three/	{eli}	I1 [BE4/012]
024 - St1:	/THREE/	{c-rpl}	R1(I1)
025 - T		8)_____	F1(R1)