DEVELOPING AND APPLYING CROSF: A NUMERIC CODE PROPOSED FOR CORPORA ANNOTATION, BASED ON HALLIDAY'S SYSTEMIC FUNCTIONAL GRAMMAR.

Marcos Pereira FEITOSA
(Universidade Federal de Minas Gerais)

ABSTRACT: *Pursuing a more efficient corpora annotation model, a numeric code to label corpora according to Halliday's Systemic Functional Grammar was proposed. This code, named "CROSF", was developed through testing several prototypes on a small parallel bilingual corpus. A case study applying CROSF's final prototype pointed out information Thematization aspects due to systemic differences between the two languages, as well as the translator's idiosyncratic choices. The numeric code facilitated searches throughout the corpus, allowing better visibility of the annotated text and also allowing searches for different combinations of Thematic choices.*

KEY-WORDS: Thematic organization; Translation Studies; Systemic Functional Linguistics; Corpus Linguistics.

RESUMO: *Visando uma anotação de corpora mais eficiente, foi proposto um modelo de anotação de corpora instrumentalizado através de um código numérico para a Rotulação com base na Gramática Sistêmico-Funcional de Halliday. Esse código, denominado "CROSF", foi desenvolvido através de sucessivos protótipos testados em um pequeno corpus paralelo bilíngüe. Um estudo de caso aplicando-se seu protótipo final apontou aspectos da Tematização da informação decorrentes de diferenças sistêmicas entre as línguas, bem como escolhas idiossincráticas da tradutora. O código numérico facilitou as buscas no corpus, permitindo melhor visibilidade do texto anotado e buscas de diferentes combinações de escolhas temáticas.*

PALAVRAS-CHAVES: Organização Temática; Estudos da Tradução, Lingüística Sistêmico-Funcional; Lingüística de Corpus.

1. Introduction

Several works have been developed in the past few decades based on Corpus Linguistics, such as researches on lexicography, clause parts, collocations and language use patterns. This paper falls into the latter type, i.e. language use patterns – more specifically, concerning Thematic organization – particularly in the field of Translation Studies. The main theoretical fundament for this paper is Michael Halliday's Systemic Functional Grammar (SFG), which started in the 1960's and has been updated in 1985, 1994 and 2004 (the latter co-written by Matthiessen), and was further developed also by several followers, among them: Eggins (1994), Thompson (1996) and Martin et al. (1997).

Pagano (2005) states that the dialogue between Translation Studies and Systemic Functional Linguistics has established a tradition of approximately four decades, although only recently has it included Corpus Linguistics. Even after this new inclusion, corpus-based researches – particularly from the 1990's – based on SFG have been heavily laborious due to the need of manually annotating the corpora according to the analysis categories requiring the analyst's intervention so that they can be correctly processed by a corpus analysis computer program such as *WordSmith Tools.* A proposal of Thematic organization analysis in parallel corpora has been put to practice among several researchers at NET (*Núcleo de Estudos da Tradução* = Translation Studies Nucleus), a group of undergraduate and graduate students of FALE (*Faculdade de Letras* = Faculty of Languages), UFMG (*Universidade Federal de Minas Gerais* = Federal University of [the state of] Minas Gerais). The first few analyses performed by Prof. Dr. Adriana Pagano and by the students themselves have shown that, in addition to taking too much time, the corpora annotation process presented flaws due to mistypes when inserting the Labels. The word *Label* is being used here according to Halliday's (1994: 24) definition, i.e. an indicator of the function each element performs in a clause. The annotation had been made through words, with the categories' names inserted fully in each Label. Such Labels ended up too long and more likely to have mistypes, which interfered in the results provided by the corpora analysis computer program. In addition, the absence of an annotation pattern among the members of NET made the dialogue among different researchers much more difficult and less productive. While there are several different computer programs for corpus annotation, none of the software available seems to categorize Themes according to their position in the sentence, Metafunction or the other

subcategories, which is why the annotation must be made manually. Also, the software used may vary from one research to another, but the annotations within the corpus could follow a similar standard for future researches using those corpora.

So as to quicken the corpora annotation process, make mistypes less likely to happen, and facilitate an interface among different researches using SFG at FALE/UFMG, a numeric code was developed as a proposal for a better means of labeling the corpora. It was the result of my Master's research, under the guidance of Professor Advisor, Dr. Adriana Pagano. This code is called CROSF: *Código de Rotulação Sistêmico-Funcional* (Systemic Functional Labeling Code). CROSF has made a better dialogue possible among those who perform researches based on SFG by means of making the same annotation system viable for all these researchers at NET and, hopefully, among other SFG researchers as well.

In order to verify the code's applicability, a pilot study was performed first, so as to build the code and, afterwards, a case study to collect and analyze data from the corpus, which was compiled of chapters 1 and 2 from the novel *The blind assassin*, written in Canadian English by Margaret Eleanor Atwood (2000) and its Brazilian Portuguese version, *O assassino cego*, translated by Léa Viveiros de Castro (Atwood 2001).

The first prototypes of CROSF and the case study also had the collaboration of Igor A. Lourenço da Silva, who was then an undergraduate student at FALE/UFMG. He helped organizing the initial tables and testing all possible settings for the *WordSmith Tools* software.

2. Describing the code

This code was developed through successive prototypes, each one of which was tested. When new flaws were discovered, a new prototype was then made. This process was repeated until a viable prototype was achieved. At the time, CROSF-14 was considered satisfactory and my Master's dissertation was concluded with that. A new model was made afterwards – CROSF-15 – simply because one category was later considered unnecessary and removed, so the final prototype is now CROSF-15, but it is basically the same code and is here provided on the Appendix at the end of this text.

CROSF uses the Theme (Textual Metafunction) as a starting point, but it

can be used for analyzes concerning any or all of the three Metafunctions. It is always compounded of a seven-digit number and may be used to label both the Theme and the Rheme, although only the Theme has been labeled for the researches carried on at NET. To make this form of annotation possible, all categories considered possible to be used in such labels described by Halliday (1994), Thompson (1996) and Martin et al. (1997) were listed and included in the code. It is important to mention that this research was conducted completely in the Portuguese language, as it was developed at a Brazilian university, so the SFG terms were all presented both in their original English names and in Portuguese translations proposed by a conjunct project of PUC-SP (Brazil) and the University of Lisbon (Portugal), which are discussed in Berber Sardinha (1999). A few other terms not found in their project were proposed by members of NET (FALE/UFMG): the Portuguese names for the Adjuncts (Interpersonal Metafunction), proposed by Cibele Bernardino, a doctorate candidate at FALE/UFMG, and a few other terms that still didn't have a name in Portuguese were translated by myself.

The code's layout is *ab cdefg*, which means that positions *a* and *b* are interrelated, while positions *cdefg* are also interrelated, so that *g* is interpreted differently depending on what value is found on *f*, which will depend on *e*, which will be interpreted differently depending on what is found on *c*. Position *d* is also understood differently depending on what is found on *c*, but no other values depend on *d*. A thorough explanation of how this works will follow now.

The code is comprised only of numbers – digits, if you may. So their positions, to avoid misunderstandings, are referred to by letters. The first digit is said to be in position *a*, the second digit is said to be in position *b*, and so on. Position *a* refers to whether this element of the clause is a Simple Theme, part of a Multiple Theme, if it is part of the Rheme or even the N-Rheme. The Rheme has also been divided here in "Simple" and "Multiple" assuming that it may have the same subcategories as the Theme. However, we have only worked with the Theme so far at NET. The N-Rheme is a category established by Fries (1994) to indicate the New information in a clause, which tends to be the last Ideational item of the Rheme. It is an important category for Thematic Progression, which might be later studied using CROSF, although this has not been made yet. In fact, studies concerning the Rheme seem to be rather scarce so far, although such studies might be performed with more emphasis in the near future.

Position *b* (the second digit on the code) indicates the position of the labeled element. If it is elliptical, number 0 (zero) is used; if it is the first or the only Theme/Rheme/etc. being labeled, *b* is 1; if it is the second part of a Multiple Theme or Rheme, the value used is 2; and so on.

Position *c* is used to indicate which of the three SFG Metafunctions it concerns: 1 for Ideational, 2 for Interpersonal, and 3 for Textual.

Position *d* is used in the Ideation Metafunction to indicate whether the labeled term is *Marked* or *Unmarked*, and in the Interpersonal Function to indicate *Modulation* or *Modalization*. As it has no possible subcategory in the Textual Metafunction, in this case it is mandatory to fill it with 0 (zero), to show that it does not apply there. In fact, every time a given category does not apply or is not regarded, a 0 should be used to fill that position. Except for position *b*, for which it actually means something, a zero means that no category applies or is regarded there.

Finally, positions *efg* correspond to a more specific analysis of the element in the clause. Therefore, if it concerns the Ideational Metafunction, for instance, and 1 is found on position *e*, the code there reads "Participant – no interpolation". More specific subcategories are found on positions *f* and *g*, all detailed on the tables for CROSF, provided in the Appendix at the end of this article.

The following example, taken from the corpus used in the research, shows how the categorizations are made:

"*A hot wind* <1111111> was blowing around my head, *the strands of my hair* <1111111> lifting and swirling in it, like ink spilled in water."

(It is important to mention here that all the examples presented on this article were taken from the corpus of the research mentioned, that is, either from Atwood (2000) or Atwood (2001). The italics are used here only to indicate what is being shown or discussed; the corpus was originally annotated in *"txt"* format, so no font formatting was used.)

The sentence in the example above shows two clauses in a relation of parataxis. Thus, two Themes are found, their Labels, coincidentally, being the same: <1111111>. All Labels must be annotated between angled brackets, so that it is made clear, when processing the corpus on a computer, that this is an annotation – not part of the original text. The

Label <1111111> is read as follows: the digit 1 on position *a* indicates it is a Simple Theme. Position *b* is 1, indicating that it is on the one and only position; were it an elliptical Theme, the value used would be 0 (zero) – even an elliptical term may be labeled with CROSF. Position *c* is 1, so it regards the Ideational Metafunction. Position *d* is 1, so it's Unmarked. Position *e* is 1, which means it is a Participant with no interpolation. Position *f* is 1, so this Participant concerns a Material Process. Position *g* is 1, which indicates that this Participant is an Actor. Therefore, the Label <1111111> reads: "Simple Theme – not elliptical – Ideational – Unmarked – Participant – concerning a Material Process – Actor. It is important to mention that, whether a Participant or a Process is being labeled, both will be described in the code.

In order to make all annotations alike and to avoid misinterpretations, it is suggested that all Labels are put right after the annotated term. This pattern is particularly useful to indicate where the annotated phrase ends. The beginning of the annotated phrase is more clearly noticed: if it is the Theme, it should be the begging of the clause. Everything else starts where the previous annotated phrase ends. So the beginning and the end of the annotated phrase should both be easily spotted.

In case the researcher does not wish to analyze all possible subcategories, any values which s/he does not wish to categorize may be filled with 0 (zero), so as to indicate that whatever that position indicates is not being observed. However, the code *must* have all seven digits. None can be omitted. Otherwise, the code may be read incorrectly by whatever computer software is used. The positions *must* be *ab cdefg*, or the code is not *one* code for all and one researcher may misinterpret another researcher's annotations. For example: supposing the researcher does not wish to state whether the item is a Theme or a Rheme, positions *ab* may be filled with "00". Position *b* with the value "0" normally reads "elliptic", but since *a* is also "0", it is understood that there is no ellipsis at all – there simply isn't a Theme or Rheme being categorized.

There is also the case of minor clauses, which present no Theme or Rheme. The Label *Absolute* may be used then, as no transitivity is observed there. No specific subcategories are supposed to be used in this case, so this Label is <5000000> throughout the corpus of this research. However, if one may wish to retrieve some transitivity from the context for a particular research, CROSF does allow the subcategorization of the item and that is not at all amiss; it may be merely unnecessary.

2.1. Multiple Labels

When the same element may comprehend more than one categorization in a clause, a Multiple Label showing both categories unified by means of a hyphen, with no spaces, may be used. For example:

"*It*'s <1111135-1111133> not my fault."

The code allows the researcher to analyze the Participant *"It"* under two different perspectives, both found on SFG: it may be understood as "Identified" or, more specifically, as "Token", making both Labels <1111133> (Simple Theme, not elliptical, Ideational, Participant with no interpolation, Identified) and <1111135> (Simple Theme, not elliptical, Ideational, Participant with no interpolation, Token) applicable. Depending on how one sets up the computer software, Multiple Themes may be found as they are, or either one of their components may be listed separately as well. Therefore, more specific and less specific searches are all viable.

3. Reading and collecting data after annotating with CROSF

Three tools available in *WordSmith Tools* suite have proved to be very favorable to the use of this numeric code: *Wordlist*, which provides lists of the occurrences of words in either alphabetical or frequency order; *Concord*, which provides collocations, showing which words co-occur with what; and *Viewer & Aligner*, which allows the researcher to view both original and translated texts simultaneously, displayed either sentence by sentence or paragraph by paragraph.

Albeit developed to show words, the tools *Wordlist* and *Concord* may both be used effectively to show the numeric Labels, as long as the software is set up to include numbers when using these tools.

Wordlist may show the Labels in order of frequency and this result may be later copied onto Microsoft Excel (or any equivalent computer program) and then reordered alphabetically, so that numbers are shown first. The researcher may simply ignore the words at the end of the list. The frequencies may then be noted. In order to view this list more easily, it may be best to simply delete the words at the end and reorder the items according to the frequency column, so that the Labels may be viewed in the order of frequency more easily.

Concord may be used to search for more or less specific Labels using a question mark (?) instead of whatever digits that are irrelevant for the search, so that the software may "accept" any values for those positions. For example, searching for <??1????>, all Ideational items will be shown; or searching for <11?????>, all non-elliptical Simple Themes will be listed.

When using *Viewer & Aligner*, the seven-digit Labels allow the visualization of the text at its side, which is not always possible when using long Labels with the full names of the categories in words.

4. The case of the ellipsis of the subject

One the difficulties found regarding different structures of each language when labeling the corpus was the ellipsis of the subject, particularly in Portuguese. SFG did foresee this possibility (*cf.* Martin et. all. 1997:29), although it is not so frequent in English. However, it is very frequent in Portuguese, even expected, so that it is, in Portuguese, *Unmarked*. Some researches on Theme dealing with the ellipsis of the subject (although retrievable from the verbal inflection) discuss about the difficulty on how to categorize the Theme in these cases. Barbara & Gouveia (2001) consider the ellipsis of the subject in the beginning of the clause to cause an Elliptical Theme, such as in the example they provide: *"Vi a casa toda."* ([I] saw the whole house.) The Theme was indicated by the authors using the symbol "Ø" to indicate the ellipsis of the Theme (Barbara & Gouveia 2001: 7). On the other hand, Ventura & Lima-Lopes (2002) consider that the Process in the beginning of the sentence is the Theme – not the elliptical subject, and they considered it to be *Marked*, which would not allow this Theme to be told apart from the cases in which the Participant is found *after* the Process. For example: in the clause *"Precisamos sair daqui!"* ("[We] need to get out of here!"), the authors have marked the word "Precisamos" as the Theme, indicating it to be the Process, hence, a *Marked* Theme.

Researchers at NET have discussed how to categorize this case of Theme, as well as whether to consider it a case of Marked or Unmarked Theme, questioning the proposal made by Ventura & Lima-Lopes; others preferred to see the elliptical subject as the Theme, annotating it as a case of Elliptical Theme, following the proposal made by Barbara & Gouveia. As this case is quite frequent in our Portuguese texts, the study of this phenomenon is quite relevant for English-Portuguese parallel corpora, once such an analysis intends to observe up to what extent the

ellipsis of the subject becomes or not a pattern in such translations. At first, I had proposed a Label as a "compromise" between both points of view: the Label *Process-Participant*, to indicate that the Theme was a co-occurrence of the Participant and the Process found on a single word, which can be understood as true in languages such as Portuguese, in which the verbal inflection often makes it unnecessary – and even undesirable, at times – to express the subject on a separate word. It is not a new category at all – it is simply a co-occurrence of the categories *Process* and *Participant* on a single word. This Label was found on CROSF-14 but has been deleted from CROSF-15 – in fact, that is the only difference between CROSF-14 and CROSF-15. Currently, the proposal of considering it a case of an ellipsis of the Theme has been adopted, since it facilitates a wider dialogue with works in other languages, such as works in Spanish, which seem to adopt this view for the same problem. So, when using CROSF-15, the case of a sentence in Portuguese beginning with the Process, with the Participant only retrievable from the verbal inflection, is categorized as a case of Elliptical Theme, the Label beginning with "10" in positions *ab*.

5. Illustrative example of the code in use

Due to length restrictions of this article, only the analysis observing the case of the former Label *Process-Participant* in Thematic position will be focused here, so as to verify in which situations the translator of the novel chosen for the corpus has opted to begin a clause with this structure. Currently, the same research would still be possible, looking for the Label: *Theme – elliptical – Ideational – Participant*. For such, all cases of elliptical subject in Thematic position may be found by using the tool *Concord* from the suite *WordSmith Tools* to search for <101?1??>.

Using the tool *Wordlist* from the same suite, a wordlist including the Labels in frequency order was found. This list was copied onto *Microsoft Excel* and, then, it was reordered alphabetically, so that the numbers, i.e. the Labels, would show first. After that, the words were deleted from the wordlist and the list was then reordered according to the column of frequency, so that a list of just the Labels in frequency order was obtained. The beginning of that list is found on Table 1 below:

	Frequency	Label
01	210	5000000
02	138	1111111
03	105	1111133
04	99	1111141
05	93	1111131
06	63	1111121
07	46	2130321
08	33	1111151
09	31	2120411
10	29	2130322
11	28	1111112
12	27	1112510
13	25	1111411
14	23	1111433
15	22	1111431
16	21	2211111
17	20	1111311
18	20	2211131
19	20	2211133
20	20	2211411

Table 1:   List of Labels, in frequency order, found in *O assassino cego*

Table 1 shows only the first 20 most frequent Labels found on the translated text. In order to make it easier to read it in this article, the line showing the 13th most frequent Label was highlighted.   The Label <1111411> reads: *Simple Theme, Ideational, Unmarked, Process-Participant, Process: Material; Participant: Actor.* According to that table, this very specific Label alone occurred 25 times on the corpus. The 20th most frequent Label <2211411> reads: *Multiple Theme, second position, Unmarked, Process-Participant, Process: Material; Participant: Actor.* This helps corroborate the possibility of considering the Theme *Process-Participant* (now understood as *Elliptical Theme – Participant*) as *Unmarked* in Portuguese, given its high frequency of occurrences in the Portuguese language corpus. In English, there were no occurrences at all of the Label <1111411> and only two occurrences of the Label <2211411>.

Since the Label *Process-Participant* did not appear frequently in the original text in English, it did not even appear on the list of the 20 most frequent Labels. The omission of the subject seems to be a recurrent practice in Portuguese, here shown by its frequent Thematic position, which helps corroborate the fact that this should be considered *Unmarked* in Portuguese. It also indicates that the translator may have been sensitive to that specificity of the Portuguese language, as she has so often chosen the ellipsis for the subject when translating that text, even without it occurring so much in the original text in English.

A few examples showing an explicit subject in English and an elliptical subject in the translation extracted from the corpus are shown below:

"I <1111111> opened the drawer, I <1111121> saw the notebooks. I <1111111> undid the crisscross of kitchen string that tied them together. I <1111121> noticed that my teeth were chattering, and that I was cold all over."

"Abri <1111411> a gaveta, vi <1111421> os cadernos. Desatei <1111421> o barbante que os prendia. Notei <1111421> que meus dentes estavam batendo e que eu estava toda gelada."

As Vasconcellos (1997) pointed out, there are cases in which a text is translated with a different Thematic structure, but that does not mean necessarily that the translation is inadequate.

6. Final considerations

The code has proved to be effective and efficient for the annotation of corpora according to Halliday's Systemic Functional Grammar and it has helped reveal differences between a text in English and its translation into Portuguese, facilitating the annotation (quicker and less inclined to have mistypes) and the visualization of the annotated corpus (much smaller Labels, as the seven-digit number is much shorter than the Label in full words). Among other specificities, when applying CROSF, it was noticeable that the ellipsis of the subject is a recurrent practice in Portuguese, so this Theme should be considered *Unmarked* in this language, i.e. it is, apparently, a usual, common form to organize the text in Portuguese, corroborating Barbara & Gouveia (2001). Although not described here, due to the limited length of the article, cases of Unmarked Themes in declarative mode with the Process placed *before* its participant, were also observed, corroborating Pontes (1987), among

other specificities of the Portuguese language. As pointed out by Baker (1992), the application of SFG onto other languages apart from English may be complex, but it is viable, with a few adaptations.

REFERENCES

ATWOOD, Margaret Eleanor. 2000. *The blind assassin.* Nova Iorque : Anchor Books.

_____. 2001. *O assassino cego.* Translated by Léa Viveiros de Castro. Rio de Janeiro: Rocco.

BAKER, Mona. 1992. *In other words*: a coursebook on translation. London: Routledge.

BARBARA, Leila; GOUVEIA, Carlos A. M. 2001. Marked or unmarked, that is NOT the question; the question is: Where's the theme?. Universidade de Lisboa/Pontifícia Universidade Católica de São Paulo. Lisboa/São Paulo. <http://lael.pucsp.br/direct/DirectPapers45.pdf>. Acessed at: 01 Dec. 2003.

BERBER SARDINHA, Tony. 1999. Representatividade de corpus. São Paulo: IX INPLA. PUC/SP. <http://lael.pucsp.br/~tony/download.htm>. Acessed at: 01 Dec. 2003.

EGGINS, Suzanne. 1994. *An introduction to Systemic Functional Linguistics.* Londres/Nova Iorque: Continuum.

FRIES, Peter H. 1994. On Theme, Rheme and discourse goals. In: COULTHARD, Malcolm. *Advances in Written Text Analysis*. London; Routledge, pp. 229-249.

HALLIDAY, M.A.K. 1994. *An introduction to Functional Grammar*. London: Edward Arnold.

HALLIDAY, M.A.K. & MATTHIESSEN, Christian M.I.M. 2004. *An introduction to Functional Grammar*. 3 ed. Londres: Edward Arnold.

MARTIN, J. R. et al. 1997. *Working with Functional Grammar*. London: Arnold.

PAGANO, Adriana. 2005. Organização temática e tradução. In: ALVES et al. *Competência em tradução: cognição e discurso.* Belo Horizonte: UFMG, pp. 247-299.

PONTES, Eunice. 1987. *O Tópico no Português do Brasil.* Campinas: Pontes Editores.

THOMPSON, Geoff. 1996. *Introducing Functional Grammar*. London:

Arnold.

VASCONCELLOS, Maria Lúcia. 1997. Can the translator play with the system, too? A study of thematic structure in some Portuguese translations". In: *Cadernos de Tradução*, n. 2. Florianópolis: Universidade Federal de Santa Catarina, pp. 149-184.

VENTURA, Carolina Siqueira Muniz; LIMA-LOPES, Rodrigo Esteves de. 2002. *O Tema: caracterização e realização em português*. John Benjamins, LAEL/Pontifícia Universidade Católica de São Paulo. <http://lael.pucsp.br/direct/DirectPapers47.pdf>. Acessed at: 01 Dec. 2003.

APPENDIX 1

CROSF-14 ("Código de Rotulação Sistêmico-Funcional" – prototype 14)

Code chart: ab cdefg

THEME/RHEME

a:
- 1: simple Theme
- 2: multiple Theme
- 3: simple Rheme
- 4: multiple Rheme
- 5: minor clause
- 6: N-Rheme

POSITION

b:
- 0: elliptic
- 1: first
- 2: second
- 3: third
- 4: fourth
- 5: fifth
- 6: sixth
- 7: seventh
- 8: eighth
- 9: ninth

FUNCTION

c:
- 1: ideational

d:
- 1: unmarked
- 2: marked

e:
- 1: participant no interpolation
- 2: participant interpolation
- 3: process
- 4: process-participant

f: **PROCESS** — g: **PARTICIPANT**

PROCESS (f)	PARTICIPANT (g)
1: material	1: actor
	2: goal
	3: recipient
	4: client
	9: range
2: mental	1: senser
	2: phenomenon
	9: range
3: relational	1: carrier
	2: attribute
	3: identified
	4: identifier
	5: token
	6: value
	9: range
4: verbal	1: sayer
	2: receiver
	3: verbiage
	4: target
	5: locution
	9: range
5: behavioral	1: behaver
	9: range
6: existential	1: existent

e:
- 5: circumstance no interpolation
- 6: circumstance interpolation

fg:
- 10: location *
- 20: extent *
- 30: manner *
- 40: cause *
- 50: contingency *
- 60: accompaniment *
- 70: role *
- 80: matter *

				90: angle *		
e	7: clause		fg	10: no interpolation		
				20: interpolation		
				1: atributivo preposto		—
	8: special cases			2: Thematized structure		1: predicated
e		f				2: equative
						3: preposed
						4: comment
						5: passive
				3: nonrepresentational pronoun		1: meteorological
						2: impersonal

* The "0" (zero) at the last position may be replaced by subcategories.

THEME/RHEME		POSITION	FUNCTION						
							1: vocative	fg	—
							2: finite (question)	fg	—
		0: elliptic					3: "wh-" element	fg	10: interrogative * / 20: exclamative*
		1: first							11: polarity
		2: second							12: probability
		3: third							13: usuality
		4: fourth							14: readiness
		5: fifth					4: modal adjunct: mood	fg	15: obligation
		6: sixth							21: time
		7: seventh							22: typicality
		8: eighth							31: obviousness
		9: ninth							32: intensity
	1: simple Theme				0: N/A **				33: degree
a	2: multiple Theme	b	c	2: interpersonal	d	1: modalization	5: metaphor	fg	—
	3: simple Theme				2: modulation				11: opinion
	4: multiple Rheme								12: admission
	5: minor clause								13: persuasion
	6: N-rheme								14: entreaty
							6: modal adjunct: comment	fg	16: presumption
									17: desirability
									21: reservation
									22: validation
									31: evaluation
									32: prediction

THEME/RHEME		POSITION		FUNCTION							
a	1: simple Theme 2: multiple Theme 3: simple Rheme 4: multiple Rheme 5: minor clause 6: N-Rheme	b	0: elliptic 1: first 2: second, 3: third 4: fourth 5: fifth 6: sixth 7: seventh 8: eighth 9: ninth	c	3: textual	d	0: (no categories)	e	1: continuative	fg	—
								e	2: structural	fg	10: coordinator *
											20: subordinator *
											31: defining relative pronoun
											32: non-defining relative pronoun
								e	3: conjunctive adjunct	fg	11: elaborating: appositive
											12: elaborating: corrective
											13: elaborating: dismissive
											14: elaborating: summative
											16: elaborating: verificative
											21: extending: additive
											22: extending: adversative
											23: extending: variative
											31: enhancing: temporal
											32: enhancing: comparative
											33: enhancing: causal
											34: enhancing: conditional
											36: enhancing: concessive
											37: enhancing: respective

* The "0" (zero) at the last position may be replaced by subcategories.
** "N/A": "Not applied".

APPENDIX 2

CROSF-15 ("Código de Rotulação Sistêmico-Funcional" – prototype 15)

Code chart: ab cdefg

	THEME/RHEME		POSITION		FUNCTION						PROCESS		PARTICIPANT
a	1: simple Theme 2: multiple Theme 3: simple Rheme 4: multiple Rheme 5: minor clause 6: N-Rheme	b	0: elliptic 1: first 2: second 3: third 4: fourth 5: fifth 6: sixth 7: seventh 8: eighth 9: ninth	c	1: ideational	d	1: unmarked 2: marked	e	1: participant no interpolation	f	1: material	g	1: actor
									2: participant interpolation				2: goal
									3: process				3: recipient
									4: [Label removed]**				4: client
													9: range
											2: mental	g	1: senser
													2: phenomenon
													9: range
											3: relational	g	1: carrier
													2: attribute
													3: identified
													4: identifier
													5: token
													6: value
													9: range
											4: verbal	g	1: sayer
													2: receiver
													3: verbiage
													4: target
													5: locution
													9: range
											5: behavioral	g	1: behaver
													9: range
											6: existential	g	1: existent
									5: circumstance no interpolation	fg	10: location *		
											20: extent *		
											30: manner *		
											40: cause *		
									6: circumstance interpolation		50: contingency *		
											60: accompaniment *		
											70: role *		
											80: matter *		
											90: angle *		
								e	7: clause	fg	10: no interpolation 20: interpolation		
								e	8: special cases	f	1: atributivo preposto		—
											2: Thematized structure		1: predicated
													2: equative
													3: preposed
													4: comment
													5: passive
											3: nonrepresentational pronoun		1: meteorological
													2: impersonal

* The "0" (zero) at the last position may be replaced by subcategories.
** Elliptic subject in Thematic position is now labeled as Theme – elliptic – ideational – participant... <10111??>.

THEME/RHEME		POSITION		FUNCTION						
a	1: simple Theme 2: multiple Theme 3: simple Theme 4: multiple Rheme 5: minor clause 6: N-rheme	b	0: elliptic 1: first 2: second 3: third 4: fourth 5: fifth 6: sixth 7: seventh 8: eighth 9: ninth	c	2: interpersonal	d	0: N/A *** 1: modalization 2: modulation	1: vocative	fg	—
								2: finite (question)	fg	—
								3: "wh-" element	fg	10: interrogative * 20: exclamative*
								4: modal adjunct: mood	fg	11: polarity
										12: probability
										13: usuality
										14: readiness
										15: obligation
										21: time
										22: typicality
										31: obviousness
										32: intensity
										33: degree
								5: metaphor	fg	—
								6: modal adjunct: comment	fg	11: opinion
										12: admission
										13: persuasion
										14: entreaty
										16: presumption
										17: desirability
										21: reservation
										22: validation
										31: evaluation
										32: prediction

THEME/RHEME		POSITION		FUNCTION						
a	1: simple Theme 2: multiple Theme 3: simple Rheme 4: multiple Rheme 5: minor clause 6: N-Rheme	b	0: elliptic 1: first 2: second, 3: third 4: fourth 5: fifth 6: sixth 7: seventh 8: eighth 9: ninth	c	3: textual	d	0: (no categories)	e	1: continuative	fg	—
								e	2: structural	fg	10: coordinator *
											20: subordinator *
											31: defining relative pronoun
											32: non-defining relative pronoun
								e	3: conjunctive adjunct	fg	11: elaborating: appositive
											12: elaborating: corrective
											13: elaborating: dismissive
											14: elaborating: summative
											16: elaborating: verificative
											21: extending: additive
											22: extending: adversative
											23: extending: variative
											31: enhancing: temporal
											32: enhancing: comparative
											33: enhancing: causal
											34: enhancing: conditional
											36: enhancing: concessive
											37: enhancing: respective

* The "0" (zero) at the last position may be replaced by subcategories.
*** "N/A": "Not applied".