Generating Abstracts from Genre Structure through Lexicogrammar: Modelling of Feature Selection and Mapping

Víctor M. Castel
Consejo Nacional de Investigaciones Científicas y Técnicas
Universidad Nacional de Cuyo
Mendoza, Argentina

Abstract: The research article abstract (RAA) has been the focus of numerous investigations within both the Systemic Functional Linguistics (SFL) and the Natural Language Processing (NLP) communities, and from both the text generation and the text analysis/parsing perspectives. Given the complexity of the object of study, however, there is still a need for extensive studies of the RAA which provide detailed descriptive generalizations on the relationship between context and language which are corpus-based, formally defined and computationally implemented. These three conditions appear to be central to any Natural Language Generation (NLG) project whose long-term goal is simultaneously to model the systemic-functional unity and diversity found in RAAs, and to use the resulting model in the development of tools for interactive rhetorical and linguistic assistance in RAA writing. This is the leading goal of the RedACTe Project one of whose theoretical-descriptive results is presented here. The basic formal mechanism used by the Cardiff Grammar for the generation of text-sentences is adapted and extended to capture systematic correlations between higher (genre and register) and lower (lexicogrammar) strata features of RAAs of the RedACTe Project's sample disciplines. Generation rules are defined, both within any one stratum and between strata, for mapping genre onto semantics and semantics onto form - and so text proper.

Key Words: NLG, NLP, SFL, Corpus Linguistics, Computational Linguistics.

Resumen: El *Abstract* del artículo de investigación científica en inglés (AbAICI) ha sido, y aún lo es, el foco de atención de numerosas investigaciones tanto desde la perspectiva de la Lingüística Sistémica Funcional (SFL) como desde la perspectiva del Procesamiento Automático de Lenguas Naturales (NLP). Dada la complejidad del objeto de estudio, sin embargo, es todavía necesario llevar a cabo estudios extensivos del AbAICI que provean generalizaciones descriptivas detalladas acerca de la relación entre contexto y lengua que estén basadas en corpus, definidas formalmente e implementadas computacionalmente. Estas tres condiciones parecen ser centrales para cualquier proyecto de Generación de Lengua Natural (NLG) cuyo objetivo a largo plazo sea simultáneamente modelizar la unidad y la diversidad sistémico-funcional encontrada en los AbAICIs, y utilizar el modelo resultante en el desarrollo de herramientas para la asistencia retórica y lingüística en la escritura interactiva de AbAICIs. Este es el objetivo central del Proyecto RedACTe, uno de cuyos resultados teórico-descriptivos se presenta aquí. El mecanismo formal básico utilizado por la Gramática de Cardiff para la generación de oraciones-texto es adaptado y extendido para expresar correlaciones sistemáticas entre rasgos de estratos superiores (género y registro) y rasgos de estratos inferiores (léxico-gramática) de AbAICIs de la muestra disciplinar del Proyecto RedACTe. Definimos reglas de generación, tanto dentro de cada estrato como entre estratos, para la proyección del género en la semántica y la semántica en la forma, y así en el texto propiamente dicho.

Palabras Clave: Generación del Lenguaje Natural, Procesamiento del Lenguaje Natural, Lingüística de Corpus, Lingüística Computacional.

INTRODUCTION

The purpose of this paper is to present the RedACTe Project's approach to the modelling of some formal and substantive properties of the generation of research article abstracts (RAAs).

RedACTe stands for *Redacción Asistida por Computadora de Textos*, which can be translated as computer-aided text writing. It is a project located at Universidad Nacional de Cuyo, Mendoza, Argentina. The main long term goal is to develop a computer application capable of assisting scientific researchers in RAA writing. To attain this goal, RedACTe draws theoretically on Systemic Functional Linguistics (SFL), particularly on the Cardiff Lexicogrammar (CLG), in order to provide a formal systemic-functional text generation oriented modelling of systematic correlations between contextual and linguistic features of the RAA. Its applied aim is to design, based on such a modelling, a system of principles for RAA writing that can be implemented into appropriate software.

The descriptive generalizations required to model the RAA are based on and validated by a sample of about 700 research articles on the following disciplines or research fields: Linguistics, Climatology, Statistics, Ecology, Arid Zones, Glaciology, Architecture, Ruminants, Science Education, Psychiatry, Waste Water Engineering, Archaeology, Sociology, Genetics, Agriculture, Internal Medicine, Robotics and Geophysics.

The resulting software is expected (i) to capture both the unity and diversity of intra- and inter-disciplinary RAA properties at both the higher and the lower strata, (ii) to adapt and implement pedagogically the necessary rhetorical and linguistic conceptualization, (iii) to assist local researchers in the writing of well-formed and appropriate RAAs, and thus (iv) to help them to overcome the comparative disadvantage that emerges when spanish-speaking researchers write RAAs to be submitted for publication in prestigious international scientific journals.

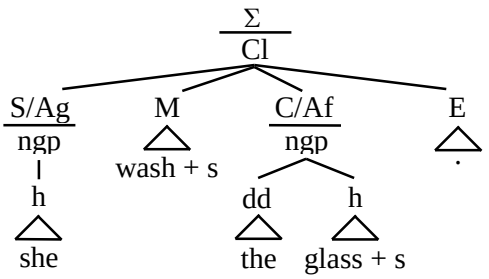
Inspired by and drawing on the CLG formal framework, I devote the rest of the paper to presenting a few theoretical-descriptive results of the RedACTe Project about correlations between genre properties, field properties and lexicogrammatical patterns.

1. The CLG Rule Formalism in a Nutshell

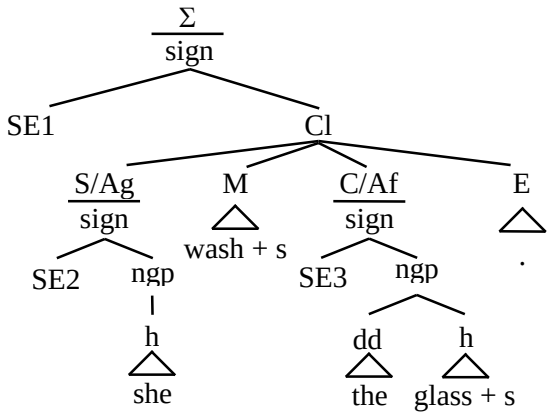
CLG can be conceived of as a set of rules capable of generating text-sentences like (1) with syntactic representations like (2) (cf. Fawcett, 2000; Fawcett et al. 1993):[1]

(1) She washes the glasses.

(2)



(3)



Key: Σ = variable ranging over genre elements; Cl = clause; S = subject; Ag = agent; ngp = nominal group; h = head; M = main verb; C = complement; Af = affected; dd = deictic determiner; E = ender; SE = selection expression; SE1 = [entity, situation, ..., present trp, ..., washing, ..., outsider sth, count sth, singular sth, ...]; SE2 = [entity, thing, outsider, recoverable thing, ..., singular tc, human tc, female tc, ...]; SE3 = [entity, thing, ..., outsider, ..., artefact, container, glass c, count cc, plural cc, ...], where 'situation', 'thing', etc. are semantic features.

Tree diagram (3) is an enhanced version of (2) used by RedACTe as a more perspicuous way to represent CLG output and the realization relationship between context of culture categories and language categories (cf. §2 below). It introduces two new nodes into the original CLG syntactic representations: a node labelled *sign* and a node labelled with a complete *selection expression* (SEn). Notice that the part of (3) headed by the topmost sign is no longer a syntactic representation but a linguistic representation, for the sign is a pair made up of a semantic structure and its associated syntactic structure.[2]

There are three types of category and three types of relationship between categories in (3): **units** (the *linguistic* unit 'sign', the *semantic* unit 'entity' heading 'SE', and the *syntactic* units 'Cl' and 'ngp'),[3] **elements** (the genre element 'Σ', and the syntactic elements 'S', 'C', 'M', 'dd', 'h', and 'E', and the participant roles 'Ag', 'Af'), and **items** ('she', 'wash', 's', 'the', 'glass', and '.'). The linguistic unit 'sign' **fills** ('___') higher level elements, a semantic unit 'SE' along with a syntactic unit ('Cl' or 'ngp') **compose** ('|') signs, elements **compose** ('|') syntactic units, and items **expound** (⏜) elements. Diagram (3) can thus be read in the following manner. The genre element 'Σ' is filled with the sign composed by the selection expression [entity, situation, ..., information, ..., present trp, ..., washing, ..., count sth, singular sth, ...], and the syntactic unit 'Cl'. This unit in turn is composed of the following elements: an 'S' conflated with an 'Ag', an 'M', a 'C' conflated with an

'Af', and an 'E'. The 'S' conflated with the 'Ag' is filled with a sign whose semantics is the selection expression [entity, thing, outsider, recoverable thing, ..., singular tc, human tc, female tc, ...], and whose form is the syntactic unit 'ngp'. This 'ngp' is composed of the element 'h' which is expounded by the lexical item *she*. Element 'M' is expounded by the lexical base 'wash' followed by the suffix 's'. The 'C' conflated with 'Af' is filled with a sign whose semantics is the selection expression [entity, thing, ..., outsider, ..., artefact, container, glass c, count cc, plural cc, ...], and whose form is the syntactic unit 'ngp'. This unit is composed of the elements 'dd' and 'h'. Element 'dd' is expounded by the lexical item *the*. Element 'h' is expounded by the lexical base 'glass' followed by the suffix 's'. Finally, element 'E' is expounded by the punctuation item ".". Let now (3) guide our presentation of the CLG rule formalism.

CLG rules involved in the generation of representations like (3) are all implications which can be represented as in (4i), read as in (4ii), and interpreted as in (4iii):

(4i) $p \Rightarrow q$

(4ii) if p, then q

(4iii) if p is true, then carry out q

where p and q are variables ranging over conditions and consequences, respectively. Condition p can be a single semantic feature, or a disjunction of semantic features, or a conjunction of semantic features.[4] Let (5) serve as examples of values for condition p:

(5i) f1 $\Rightarrow q$

(5ii) f2 / f3 / f4 $\Rightarrow q$

(5iii) f5 & f6 & f7 $\Rightarrow q$

(5iv) f8 / f9 / (f10 & f11) $\Rightarrow q$

(5v) f12 & (f13 / f14) & f15 $\Rightarrow q$

Consequence q can be a(n) (conjunction of) operation(s), and/or a(n) (conjunction of) implication(s) like (4i). Let (6) serve as examples of values for consequence q:

(6i) $p \Rightarrow$ #% f1 / #% f2 [sp##] / #% f3 [rr##].

(6ii) $p \Rightarrow$ (#% f5 / #% f6) & (#% f7 / #% f8).

(6iii) $p \Rightarrow$ **sm_ps prefer** sn## <#% f1 & #% f2>.

(6iv) $p \Rightarrow$ *sign* **fills** *element*.

(6v) $p \Rightarrow$ *element* or *unit* **@** *place in unit*.

(6vi) $p \Rightarrow$ *element* **<** *item*.

(6vii) $p \Rightarrow$ *prl* **by** *element*.

(6viii) $p \Rightarrow$ **for** *prl* **re_enter_at** entity.

(6ix) $p \Rightarrow$ **for** *prl* **prefer** sn## <#% f4 & #% f5>.

(6x) $p \Rightarrow$ (f6 / f7) & (f8 / f9) $\Rightarrow q$

Key: 'f1', 'f2', etc. stand for semantic features; '/' and '&' stand for the logical operators 'or' and 'and', respectively; '#' is a variable ranging over 1 to 100, so that '#%' is the probability associated with the feature it is attached to; ($\#_1 + \#_n + \#_{n+1} = 100$); 'sp' and 'rr' abbreviate 'same pass preference resetting rule' and 'realization rule', respectively'; 'sn' abbreviates 'system network rule'; '##' is a variable ranging over rule numbers; the expressions in bold define the operators of the operations to be carried out on a given structure; the expressions in italics are variables ranging over categories; *prl* ranges over participant roles; 'entity' is the initial value for *p* and its presence triggers system network rule application.

How is the truth value of *p* determined and what does "carry out *q*" mean (cf. (4iii))? The answers to these two questions will show, respectively, the essential similarities and differences among all CLG rule subtypes.

CLG rules are organized into two components: the semantic component and the form component. The semantic component contains two different subtypes complying with the implication (4i): *system network rules* (SNRs), and *same pass preference resetting rules* (SPRs). The form component contains three subtypes of *realization rule* (RR) which also comply with (4i): realization rules proper and graphological rules.[5] The set of all CLG rules allow for the generation of linguistic representations like (3). Let us now elaborate on the distinctive traits of each subtype.

SNRs and SPRs jointly define semantic representations (= selection expressions), i.e. sets of experiential, interpersonal, and textual features. The task of an SNR is to construct a selection expression by introducing one or more semantic features into a representation (carry out *q*) whenever its entry condition *p* is met, i.e. is true. Notice that SNRs (6i-ii), as oppossed to SPRs (6iii) and RRs (6iv-ix), do not have explicit operators defining the relevant operations in

consequence *q*. This absence in (6i) is to be interpreted as a default operator indicating to introduce one of the disjuncts into the semantic structure being constructed; the absence in (6ii) is to be interpreted as an instruction to introduce, first, one of the disjuncts of the leftmost conjunct, and, second, one of the disjuncts of the rightmost conjunct.[6]

As illustrated in (6i), a feature in consequence *q* of an SNR can have a reference to an SPR or an RR associated with it, or both an SPR and an RR reference. Whenever a feature is chosen which has an SPR reference associated with it, this SPR must be applied at once. This is not the case of RR references; these references serve the purpose of allowing the system to store them so that they are applied once a given pass through the network is completed, i.e. after all SNRs and SPRs have been applied, and so a selection expression is constructed.

The task of an SPR is to *modify* the initial probabilities associated with features in SNRs in a given pass through the network.[7] Example (6iii) illustrates an SPR which instructs (the user of) the system to go to rule sn## and alter the probabilities associated with its features according to the values specified in (6iii); **sm_ps prefer** means: for the current pass you are working through, prefer this new probability assignment.

The result of applying cyclically the set of SNRs and the relevant SPRs is a selection expression examples of which are SE1, SE2 and SE3 in (3).

Rules (6iv-ix) are examples of RRs. The task of these rules is to define form representations, i.e. tree structures which account for syntactic, lexical, and punctuational (or intonational) properties of text-sentences realizing a given semantic representation.

Diagram (3) illustrates essential RR operations. The topmost sign in (3) is inserted by an instance of rule (6iv) so that it fills the higher level element 'Σ'. Both a semantic unit, which is the result of applying the set of SNRs, and a syntactic unit are inserted by an instance of rule (6v) so that they compose the sign filling 'Σ'. It is also by instances of rule (6v) that the elements 'S', 'M', 'C', and 'M' get located at the relevant places in 'Cl'. Then, the participant roles 'Ag' and 'Af' are conflated with the elements 'S' and 'C', respectively, by instances of rule (6vii). After this, instances of rule (6vi) expound 'M' and 'E' as the lexical base 'wash' followed by the suffix 's', and the punctuation item '.', respectively. Crucially, thus, the syntactic unit 'Cl', and therefore its component elements with their expounding items, *realize* SE1. This concludes the RR application cycle governed by the selection expression SE1. Instances of rule (6viii)[8] allow for

reentering the set of SNRs so that new signs are built up for filling 'S/Ag' and 'C/Af', and thus the complete representation (3) is generated.

All CLG rules, SNRs, SPRs, and RRs, share the property of assigning condition *p* three possible classes of value: a single feature, a feature disjunction, or a feature conjunction. They also share the domain type upon which the truth value of condition *p* is determined, namely: (sub)sets of features composing a selection expression.[9] CLG rules differ in the operations they perform (the value assigned to consequence *q*): SNRs construct selection expressions by inserting semantic features into structures; SPRs alter SNRs by modifying the probabilities assigned to SNR features; RRs perform various other types of operation (filling, composition and exponence, among others).

2. The RAA Genre Structure

One aspect that genre theory must address is the problem of defining where what can be in what kind of relationship with what under what kind of circumstances with what kind of interpersonal stance on the part of the writer. The thematic position chosen for "where" in this indirect question points to the central role played by the text **location** of the content to be communicated. This is known in SFL as the text *generic structure potential* (GSP; Hasan 1984, 1996: 53; Martin 1992: 550).

The RedACTe Project draws on genre theory to account for the distribution of lexicogrammatical patterns in the RAA (Bazerman, 1988; Swales, 1990; Weissberg & Buker, 1990), i.e. to define where exactly in the RAA what can or must be in what kind of relationship with what under what kind of circumstances with what kind of evaluative stance on the part of the writer.

I am deliberately using new content seekers - or, in other frameworks, wh-words or wh-headed phrases - to refer, in very abstract and neutral terms, to both the general and the domain-specific knowledge which is communicated at appropriate locations of the RAA.

A careful reading of Hlavacka (2004), which is an example of an SFL multi-stratal predetermination model,[10] suggests that the distribution of lexicogrammatical patterns (clause types, process types, lexical exponence, etc.) in the RAA is predetermined, at least in part, by the distribution of objects (or events) and relationships between objects (or events) as these are conceptualized in

scientific research in general and in each of the different scientific fields of the RedACTe Project's sample disciplines.

This experiential approach to the study of the RAA soon encounters the fundamental problem for any natural language generation project which is the need to find a definition of what we might call the *field structure potential* (FSP), and of the systematic correlations between this potential and the lexicogrammatical potential (LGP).

The main difficulty we face in RedACTe in relation to the definition of the FSP is that we are not experts on any of the sample fields, except perhaps partially on a few subfields of Linguistics. What we do in practice is either interact with researchers who indeed are experts on a given field so that they provide us with the relevant field information, or simply infer some aspects of the FSP as it is reflected by the lexicogrammar that we, as linguists, are trying to contribute to define.

This experiential approach to the definition of the RAA GSP is currently being complemented by team work inspired partly in Boccia (2001)'s interpersonal approach to the study of the RA Introduction, and partly in Rezzano (1999, 2003)'s contribution to epistemic modality in RAs. So, for example, some aspects of Boccia (2001)'s categories 'Justificación I', 'Justificación II', and 'Ofrecimiento', which are defined to capture interpersonal properties whose lexicogrammatical reflection shows a text structuring which roughly corresponds to Swales (1990)'s categories 'Establishing the territory', 'Establishing the niche', and 'Occupying the niche', respectively, are now being incorporated into Hlavacka (2004)'s definition of the RAA Introduction so that both the interpersonal and the experiential stances are encompassed by the RedACTe RAA GSP.

Thus, RedACTe conceptualizes the RAA GSP as a definition of the class of structured locations which function as placeholders where certain pairs (and not others) of register representations and linguistic representations (i.e. signs) are situated. Put in slightly different terms, terminal genre categories are well-delimited text locations where the writer is expected to communicate a certain experiential content by adopting a certain interpersonal stance on such a content and presenting it as information which facilitates the reader's comprehension of the writer's conception and evaluation of the domain specificity.[11] See Figure 4 below for examples of the highly constrained types of text-sentence which typically fill the RAA genre category [wth_rch_annc] (and only this RAA genre category, i.e. such text-sentence types could not fill

any of the other terminal genre categories defined by the Genre Selection Rules in Figure 2 below) .

As shown in §2.3 below, I resort to the **filling** relationship which in CLG is used to relate syntactic units such as 'Cl', 'ngp', etc. to elements such as 'Ag', 'Af', etc., and extend it to formalize the relationship between linguistc signs and terminal genre categories such as [wth_ttl], [wth_rch_annc], etc. (see the Introduction and cf. Figures 2-3). The intuition captured by this formalization is precisely that certain virtual text locations are filled with certain signs whose semantics is associated with certain register properties. RedACTe is progressively elaborating on these three occurrences of "certain", and defining the systematic correlations among three types of potential structure: GSP, *register structure potential* (RSP; especially, FSP), and LGP.

2.1. A Formal Definition of the RAA GSP

The long term goal of RedACTe is to define a text grammar capable of generating texts like the RAA (7) in Figure 1.[12] In the current phase of development, RedACTe is exploring the consequences of the grammar organization proposed in Castel (2004; 2005b), which assigns the RAA GSP the central role of initiating the text generation process.

The Genre Selection and Intra-stratal Genre Projection Rules in Figure 2 formalize both essential aspects of Hlavacka (2004)'s description of the RAA genre structure in general, and Miret (in preparation)'s description of the genre feature [wth_abs_ccl_rmd] in particular. The task of these rules, along with the inter-stratal genre projection rules, is to define genre feature structures like (8) in Figure 3.

From a formal point of view, notice that these rules are all implications of the form (6i-iii, ix). From a substantive point of view, the joint effect of the three subtypes of genre rules in Figure 2 is to allow for the explicit enumeration of a set of feature structures which define a class of texts according to their generic organization, i.e. they define the RRA GSP.

(7)

U0 Sex Differences in Brain Morphology in Schizophrenia

U1 Peg Nopoulos, Michael Flaum and Nancy C. Andreasen

U2 The current literature on sex differences in schizophrenia with regard to structural brain abnormalities is inconsistent. **U3** Several studies have suggested that male and female patients may differ in severity of brain abnormalities. **U4** Efforts to explore this issue have been hindered by small study groups, unbalanced groups (i.e., those with many more men than women), or both. **U5** The relatively smaller number of female schizophrenic patients in most studies may have made it more difficult to detect differences between patients and comparison subjects. **U6** This study was designed to evaluate brain morphology in a carefully selected group of patients with schizophrenia and healthy comparison subjects who were balanced by sex. **U7** Eighty patients (40 male and 40 female) and 80 healthy volunteers matched by sex and age were studied. **U8** Magnetic resonance imaging scans were analyzed with the use of an automated method that yields volumes of major brain regions. **U9** There was a significant sex-by-diagnosis interaction for ventricular volume, with male patients having significantly larger ventricles than male comparison subjects, but female patients showing no significant enlargement in comparison with healthy female subjects. **U10** Although the overall distribution of structural brain differences was very similar in the male and female patients, the male patients had a greater number of significant abnormalities than the female patients. **U11** These findings indicate that male and female patients with schizophrenia have the same pattern of structural brain abnormalities, but male patients appear to manifest greater severity, especially with regard to ventricular enlargement. *American Journal of Psychiatry* 154: 1648-1654, 1997.

Figure 1. Example of RAA.

Genre selection

cc_s1: ctx_ctr_entity ⇒ (md_sc [cc_sp1] / clm / sttt [cc_sp2] / lng [cc_sp2] / ww [cc_sp3] / …) &
(0% cnf_ppr / 100% rch_art / 0% rvw_art / 0% doc_diss / …).

cc_s2: md_sc ⇒ psych / int_md / srg.

cc_s3: rch_art ⇒ (100% wth_ttl [cc_lp2] / 0% wtht_ttl) & (100% wth_ath [cc_lp3] / 0% anonymous) &
(90% wth_abs / 10% wtht_abs) & (95% wth_art_intro / 5% wtht_art_intro) &
(60% wth_art_mth / 40% wtht_art_mth) & (95% wth_art_rsts / 15% wtht_art_rsts) &
(70% wth_art_ccl / 30% wtht_art_ccl) & (98% wth_ref / 2% wtht_ref).

cc_s4: wth_abs ⇒ (94.65% wth_abs_intro / 5.35% wtht_abs_intro) & (0% wth_st_ar / 100% wtht_st_ar) &
(52.94% wth_abs_mth / 47.06% wtht_abs_mth) &
(72.19% wth_abs_rsts / 27.81% wtht_abs_rsts) &
(0% wth_ills / 100% wtht_ills) & (73.26% wth_abs_ccl_rmd / 26.74% wtht_abs_ccl_rmd).

cc_s5: wth_abs_intro ⇒ (44.38% wth_prl_gnr [cc_lp4] / 55.62% wtht_prl_gnr) &
(19.78% wth_prv_st_lm [cc_lp4] / 80.22 wtht_prv_st_lm) &
(73.26% wth_rch_annc [cc_lp4] / 26.74% wtht_rch_annc).

… (Rules defining [wth_st_ar], [wth_abs_mth], [wth_abs_rsts], and [wth_ills].)

cc_s6: wth_abs_ccl_rmd ⇒ (98% wth_abs_ccl / 2% wtht_abs_ccl [cc_sp4]) & (25% wth_rmd / 75% wtht_rmd) &
(37% wth_rch_evl [cc_lp4] / 63% wtht_rch_evl).

cc_s7: wth_rmd ⇒ (63% wth_rmd_prp [cc_lp4] / 37% wtht_rmd_prp [cc_sp5]) &
(37% wth_lm_stmt [cc_lp4] / 63% wtht_lm_stmt).

… (Rules defining [wth_art_intro], etc.)

Intra-stratal Genre Projections

cc_sp1: md_sc ⇒ sm_ps prefer cc_s3 <100% wth_abs & 100% wth_art_intro & 100% wth_mth &
100% wth_rsts, 100% wth_ccl &100% wth_ref>,
cc_s4 <100% wth_abs_intro & 100% wth_abs_mth & 100% wth_abs_rsts &
100% wth_abs_ccl_rmd>.

cc_sp2: sttt / lng ⇒ sm_ps prefer cc_s4 <20% wth_ills & 80% wtht_ills>.

cc_sp3: ww / ar_zns ⇒ sm_ps prefer cc_s4 <20% wth_st_ar & 80% wtht_st_ar>.

cc_sp4: wtht_abs_ccl ⇒ sm_ps prefer cc_s6 <100% wth_rmd & 0% wtht_rmd>.

cc_sp5: wtht_rmd_prp ⇒ sm_ps prefer cc_s7 <100% wth_lm_stmt & 0% wtht_lm_stmt>.

… (Other intra-stratal genre projection rules.)

Inter-stratal Genre Projections

cc_lp1: *tml_gnr_ft* ⇒ sign fills *tml_gnr_ft*, lgc_entity @ 1, entity @ 2,
for *tml_gnr_ft* enter_at lgc_entity, for *tml_gnr_ft* enter_at entity.
cc_lp2: wth_ttl ⇒ for lgc_entity prefer lf1 <96% object, 4% event>, apply ttl_dalg1.
cc_lp3: wth_ath ⇒ for lgc_entity prefer object, apply ath_dalg1.
cc_lp4: wth_rch_annc ⇒ for lgc_entity prefer event, apply annc_dalg1.
… (Other inter-stratal genre projection rules.)

Key: cc_s = context of culture selection rule; ctx = context; ctr = culture; md_sc = medical science; clm = climatology; sttt = statistics; lng = linguistics; ww = waste water engineering; cnf_ppr = conference paper; rch = research; art = article; rvw= review; doc_diss = doctoral dissertation; psych = psychiatry; int_md = internal medicine; srg = surgery; ar_zns = arid zones; wth = with; wtht = without; ttl = title; ath = author; abs = abstract; intro = introduction; mth = method; rsts = results; dscs = discussion; ccl = conclusions; prl = preliminary; gnr = generalization; prv = previous; st = studies; lm = limitations; annc = announcement; rmd = recommendations; evl = evaluation; rstmt = restatement; int = interpretation; prp = proper; stmt = statement; cc_sp = context of culture same pass preference resetting rule; cc_lp = context of culture lower pass; tml_gnr_ft is a variable ranging over terminal genre feature; lgc = logic

Figure 2. RedACTe Genre Rules.

(8)

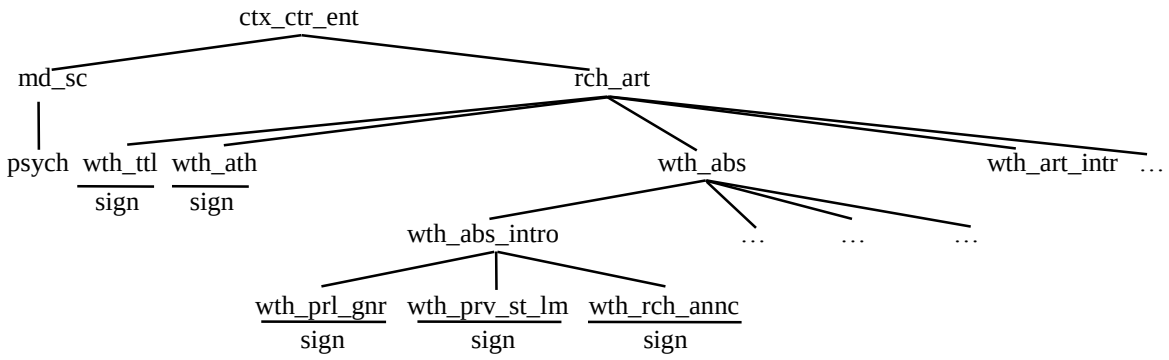


Figure 3. Genre structure instance with terminal features filled with linguistic signs.

Genre Selection Rules cc_s1-7 serve the analogous purpose of CLG SNRs of explicitly defining feature structures (cf. (6i-ii)), except for the nature of the features. Intra-stratal Genre Projection Rules cc_sp1-5 preset the RAA GSP, i.e. they constrain the class of genre feature structures on the basis of other genre features. They behave as CLG SPRs in that they also act as implicit system networks (cf. (6iii)). By 'projection' I mean the capacity that some rules have to affect properties of other rules independently of whether these belong in the same stratum or in lower strata than the former. The restriction is that the affected rules apply after the affecting rules. The function of Intra-stratal Genre Projection Rules is to map genre feature preferences onto other genre features.

The function of Inter-stratal Genre Projection Rules is to map genre feature preferences onto Logical Form (LF) features. Rules cc_lp2-4 predetermine the selection of LF features on the

basis of genre features. Thus, they define a realization relationship between the context of culture stratum, and (the field properties of) the context of situation and the language strata (cf. (6ix)).

2.2. Generating RAA Genre Structure Instances

How do we generate the feature structure (8) in Figure 3? My answer to this question will ignore the important issue of identifying and defining the factors that are relevant for modelling how we decide what to choose at a given choice point in a system network.[13] I will simply assume that it is the user of the generation system or the generation system itself, depending on whether it is set in the interactive or the random generation mode, that makes the relevant choices when these are available at whatever stratum a decision must be made. I will also assume that the user has decided to write a RA on Psychiatry.

So the user asks the system to help him write the RAA by clicking on an appropriate menu option which initiates the generation process by allowing him to pick up Medical Sciences (cc_s1) and Psychiatry (cc_s2) from among the set of available disciplines.[14] The user could not choose a text type other than the RA, for the features representing other types have 0% probability assigned to them.[15] The user's choice of Medical Sciences triggers the immediate and automatic application of rule cc_sp1, whose effect is to replace the default probabilities associated with certain features of rules cc_s3 and cc_s4 by the values specified in cc_sp1. The system driven choice of [rch_art] triggers the application of rule cc_s3 which, in its initial state, offers the user to decide among various genre configurations for the RA he is planning to write. But notice that some of the probabilities in rule cc_s3 have been altered by cc_sp1 with the net effect that such alternative configurations will not be available to the user, for the rule has been transformed into a conjunction of obligatory features. Thus, the system introduces into the genre structure under construction the subtree headed by [rch_art] in (8). The generalization captured by the joint effect of rules cc_sp1 and cc_s3 is that Psychiatry RAs are organized in such a way that the user will be assisted in the writing of the following "sections": [title], [author], [abstract], [introduction], [method], [results], [conclusions] and [references].[16] Similarly, rules cc_sp1 and cc_s4 jointly stipulate that the Psychiatry RAA will have the genre structure shown by the subtree headed by [wth_abs] in (8); intuitively, the user will be assisted in producing texts for the sections [introduction], [method], [results] and [conclusions]. The user is then offered to choose

one of the feature configurations offered by rule cc_s5 which I will assume is the subtree headed by [wth_abs_intro] in genre structure (8).

This is a good place to draw a distinction between genre structure features and section titles heading specific text locations. Genre features are used to define text organization properties at a prerealization level, and as such they are not section titles. The realization or not of genre features as section titles is a function of the lexicogrammatical resource as it is predetermined by genre projections. It is interesting to remark at this point that the medical science RAA is so strongly conventionalized, that the leading journals explicitly assign section titles to it which authors must fill with text. However, journals of other disciplines do not predefine section titles for the RAA organization, although some conceptual organization may be recommended in style sheets. RedACTe can easily account for this variation by defining projection rules so that, depending on the discipline and journal, RAA genre features can or must be linguistically realized, i.e. the relevant text location can or must be given a title. Thus, for example, depending on the medical science journal, the RAA genre feature [introduction] can be given the title "Background" or "Objectives" or both; the RA genre feature [conclusions] must be given the title "Discussion", whereas the RAA genre feature [conclusions] can be given the title "Conclusions" or "Interpretation".[17]

Had the user chosen to write an RA on Statistics or Linguistics, the system would have asked him to choose between writing an RAA with or without a section for [illustration], an option which none of the other disciplines has available, for they simply do no allow for such a section (cf. cc_sp2 and cc_s4). Similarly, only if the user had chosen to write an RA on Waste Water Engineering or Arid Zones, would he have been then offered to choose between an RAA with or without a section for [area of study] (cf. cc_sp3 and cc_s4). The function of an intra-stratal projection rule like cc_sp4, in its interplay with the genre selecton rule cc_s6, is to make sure that if the user happens to choose to have a section [conclusion - recommendation] without a subsection [conclusions], then he will be forced to have a subsection [recommendations]. An analogous function is served by rule cc_sp5 whenever the option offered by cc_s7 of not having a subsection [recommendations proper] is taken by the user, i.e. the subsection [statement of limitations] is made obligatory.

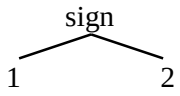
2.3. Mapping Genre Structure onto Signs

The general function of inter-stratal genre projection rules is to formally account for the realization relationship between genre features and linguistic signs, i.e. to define systematic correlations between the GSP and the LGP. Thus, they map genre features onto lexicogrammatical features in the sense that the former predetermine the latter.

Rules cc_lp1-4 look very much like CLG RRs. In fact, they are formally identical, for they comply, in general, with the implication (4iii) and, in particular, with conditions (5i-ii) and consequences (6iv-v, viii-ix). They differ, however, in the following substantive aspects related to the way consequence *q* is defined.

As pointed out in the Introduction, RedACTe enhances CLG by incorporating a new category with the status of 'unit', namely, the (linguistic) *sign*. Now it is the unit 'sign', not the unit 'cl' or 'ngp, etc., which fills higher level elements, either terminal genre features ('Σ') or elements such as 'Ag', 'Af', etc. (cf. operation **sign fills *tml_gnr_ft*** in cc_lp1). The sign is a unit with two places, in the sense of the CLG theory of unit potential structures, as represented in (9):[18]

(9)



The sign is composed of the semantic unit 'entity' at place 1 (cf. operation **entity @ 1** in cc_lp1), and a syntactic unit ('Cl', 'ngp', etc.) at place 2. This syntactic unit is treated as if it were an element, for an operation locating it at 2 in (9) is needed (**Cl @ 2**, etc.). This composing operation is defined by the relevant RRs associated with features of the selection expression headed by [entity]. The sign is thus a linguistic structure made up of a semantic structure and an associated syntactic structure which realizes it.

The fifth operation in cc_lp1, **for *tml_gnr_ft* enter_at entity**, stipulates that, for each terminal genre feature, CLG SNRs apply to define a selection expression at place 1 of the sign realizing the terminal genre feature. Notice that this operation *initiates* the system nework traversal and that, in this respect, it differs minimally from (6viii) in that this rule subtype allows for the reapplication of SNRs.

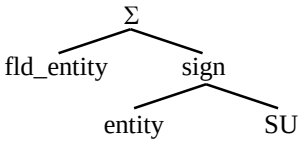
Inter-stratal genre projection rules cc_lp2-4 comply with the implication subtype (6ix). They are LF predetermination rules for they enforce preferences of terminal genre elements on LF Rules (LFR) features. This effect is attained by altering the system initial state feature probability settings, in most cases by assigning a 100% value. These genre projection rules differ from CLG preference altering RRs in the nature of the features triggering them: genre features for the former and LF features for the latter. Inter-stratal Genre Projection Rules also trigger the application of algorithms guiding LF construction like, for example, the Research Announcement LF Selection Algorithm (see Figure 5 below).

3. The RAA Register Properties

Since projection rules like cc_lp2-4 explicitly correlate genre features with LF features, formalizing the hypothesis that it is genre that predetermines field properties and, therefore, indirectly, lexicogrammatical patterns, one can legitimately ask how register fits the RedACTe Project's conception declared in §2 above, namely, that it is **both** register **and** language properties which fill terminal genre features.

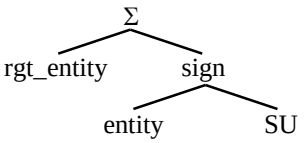
Castel (2004) proposes a text grammar which (a) treats field properties as belonging to a stratum lower than the genre stratum but higher than the lexicogrammatical stratum, and (b) defines field preferences that alter semantic feature probabilities in SNRs. This text grammar associates, with each terminal genre feature, a pair made up of a selection expression headed by the feature [field_entity], and a linguistic structure as in (10), where SU is a variable ranging over syntactic units:

(10)



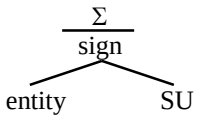
By elaborating on this higher stratum system network rule treatment of field properties, RedACTe could add tenor and mode systems so that (11) would be generated instead of (10), where [register_entity] heads a selection expression composed of field, tenor and mode features:

(11)



In CLG, however, tenor and mode properties are treated not as part of a separate stratum from the lexicogrammar but as part of tenor and mode systems which interact with semantic systems **within** SNRs. Tenor and mode features are thus "embedded" in various and complex ways within paths of the feature structure headed by [entity] in (12):

(12)



Since the rationale underlying the RedACTe Project includes taking full advantage of CLG as an existing systemic functional text-sentence generation resource, it would not be wise to rewrite CLG so that it can generate representations like (11) instead of (12). Furthermore, as far as tenor and mode features are concerned, there seems to be no essential difference between one treatment and the other as long as these two subtypes of register feature are entered first in the system network traversal so that preferences for semantic features based upon them can be defined. Then, RedACTe adopts the CLG approach to the treatment of tenor and mode features as a dimension within the scope of SNRs and thus defines representations like (12) to account for them.

It is not obvious, however, how and where CLG would account for field properties. I will assume that this register variable can be specified, as suggested by the diagram (21) in §3.2 below, at a level higher than the lexicogrammar but lower than the genre grammar, as some sort of LF along the lines suggested by Fawcett (in preparation).

3.1. Genre Driven Field Structure Potential

Every (fragment of) text, whether it is a title, or a paragraph initial clause, or a complete paragraph, or an RAA, or a complete RA, etc. refers to a conceptual configuration (CC) built upon two fundamental bricks, namely: objects and events. Both objects and events can have properties associated with them, basically, qualities and quantities. They can also be complex,

i.e. be defined as a relationship of some sort with other objects and/or events. An event is a relationship between two objects and/or events which is temporally contextualized.[19] I use the term *relationship between* to refer, with maximum generality, to any type of predicate which turns out to be necessary to include in the knowledge base (KB) which is supposed to underlie the construction of the logical expressions which constitute the initial input to the generation process.[20]

I assume that this connection between the text and the CC is the referential function of language.[21] The CC exists in the Performer's mind (writer or speaker), and therefore it is a representation of his knowledge of (an aspect of) the world. This paper draws on the concept of CC, specifically on the following hypotheses which are explored by París & Castel (2005) in an attempt to account for inter-clausal connections in the RAA as these relate to genre structure:

H1: Every clause or group refers to a CC.

H2: A CC is produced in the context of the KB. A CC presupposes the existence of other CCs which contextualize it. The connections with the presupposed CC is not always explicit.

H3: A CC opens a potential for the creation of other CCs. A CC creates expectations on other CCs which deploy it. The connection potential with forthcoming CCs is, in principle, unlimited.

H4: A CC can be defined as a logical expression which is interpretable in the KB.

H5: A CC is a construct which belongs in a higher stratum than the strata of the grammar of genre, the grammar of rhetorical relations, and the lexicogrammar.

Hypotheses H1, H3, and H3 suggest that a given clause of a given text should have some sort of connection with other clauses of the text both above and below it. The study of what can be called the *inter-clausal connection potential* has been, and still is, the main challenge to any text theory interested in discovering underlying textual organization patterns (see, for example, Mann et al., 1992).

Let us now consider the sample RAA (7). What are the objects and relationships between objects referred to by the component units U0-U6 of (7), i.e. what are its CCs?[22] The following list contains essential, though by no means all, aspects of such CCs and of the "process" that, I claim, underlies its creation.

(13i) There is an object **ob0**, referred to by **U0**, which can be conceptualized as a *relationship*, not yet specified in detail, *between sex differences*, on one hand, and *brain morphology in schizophrenia*, on the other. Notice that the second term of this relationship, "brain morphology in schizophrenia" is in turn a relationship between two terms, "brain morphology" and "schizophrenia".

(13ii) There is an object **ob1**, *the researchers - writers*, referred to by **U1**. In their role as researchers, **ob1** occupies a priviliged position in the scientific research activity as compared with all other objects composing it; in fact, they possess control over all types of decision involved in the research activity itself.

(13iii) There is a class of objects **ob3**, *other researchers - writers*, which represents researchers other than the researchers referred to by **U1**. There are no explicit references in (7) to objects of this class, but we know by KB that if there are references to other studies, then there are researchers who carried them out (cf. (13v)). The RedACTe Project's discipline sample contains RAA instances with explicit references to **ob3**.

(13iv) There is a class of objects **ob4**, *the research* which is designed and carried out to answer a scientific question. References to instances of this class are observed in **U2-U6**.

(13v) There is a relationship **define (ob3, ob4)**: an instance of **ob3** defines an instance of **ob4** so that **ob3** assigns **ob4** the property of having objectives, hypotheses and methodology. We know by KB that the function of the methodology is to guarantee the attainment of the objectives and the testing of the hypotheses. It is thus obvious that the objectives, hypotheses and methodology defined for **ob4** by **ob3** are the product of **ob3**'s decisions.

(13vi) **ob0** is more delicately conceptualized by **U2-U5** as a question (**question**) posed by other researchers (**ob3**) on whether there are differences (**differ**) in the brain morphology of schizophrenic patients (**ob5**) which are correlated with these patients' sex differences (**ob6**): **question (ob3, differ (ob5, ob6))**.

(13vii) We know by KB that, to answer the question posed in (13vi), it is necessary to observe **ob2**, more exactly, the *relationship* between **ob5** and **ob6**. Then, the research **ob4** underlying (7) forsees a relationship between **ob3** and the correlations between **ob5** and **ob6**: **observe (ob3, correlate (ob5, ob6))**. There are no explicit references to this relationship in (7), but we know by

KB that the answer to a scientific question requires observing the object of study in accordance with a given methodology.

(13viii) There is a class of objects **ob7** which are answers to the question posed in (13vi) based on the observations in (13vii). It is this class - the results proposed by other researchers - which **ob1** evaluates as inconsistent (cf. **U2**). This type of answers is classified in KB as *results*.

(13ix) The researchers **ob1** also pose to themselves the question in (13vi). To answer it, they also define a research activity **ob4** ((13v)) whose central objective is precisely to find an answer to such a question. To attain it, they must also observe the object of study **ob0** ((13vii)), and thus find another instance of **ob7** which is more reliable than the existing ones. This surpassing status of the new instance of **ob7** is due to the incorporation by **ob1** (and therefore by **ob4**) of properties relative to the study group: now the group will have the same proportion of men and women.

(13x) There is an object **ob8** which is the *article* in which the researchers **ob1** communicate the results **ob7** to potential readers.[23] There are no explicit references to **ob8** in (7), but there are other RAAs in the RedACTe text sample in which references to **ob8** are explicit.

(13xi) It follows from (13x) that there is an object **ob9** which is the potential reader of **ob8**.

(13xii) It also follows from (13x) that there exists a relationship between the researchers (**ob1**), the results (**ob7**), the article (**ob8**), and its potential readers (**ob9**). The researchers communicate the results of their study to potential readers in an article: **communicate (ob1, ob7, ob8, ob9)**.

3.2. Towards a Research Announcement LF

As stated in §2, RedACTe is expected to define where exactly in the RAA what can or must be in what kind of relationship with what under what kind of circumstances with what kind of evaluative stance on the part of the writer. This question assigns a predetermination role to the GSP over the RSP, particularly, over the FSP, in the sense that only certain CCs are preferred on the basis of the terminal genre categories within which they are constructed.

Descriptive work carried out within the RedACTe Project (Car, 2000; Castel, 2001, Castel & Diblasi, 1999; Boccia, 2001; Hlavacka, 2004; Miret, 2002, in preparation; Rezzano, 1999, 2003) shows conclusively that certain lexicogrammatical patterns (and not others) systematically occupy certain text locations (and not others). This regular distribution allows for the

identification of objects and relationships between objects which are refered to in each specific text location. The data resulting from these descriptive observations are then interpreted as indicating that terminal genre categories predetermine certain CCs which in turn predetermine the lexicogrammatical organization capable of expressing them. Put differently, the FSP appears to be defined in accordance with preferences enforced by terminal genre categories. The examples in Figure 4, which are extracted from the RedACTe Project's sample of RAAs on Psychiatry, Waste Water Engineering, Linguistics, Statistics, and Science Education, are all instances of signs filling the terminal genre category [research announcement] (cf. (8)).

{¬ ob1, ob4, ¬ ob8, …} (cf. 20vi)

(14i) In the present study, combined data from the two trials were analyzed.

(14ii) This study was designed to evaluate brain morphology in a carefully selected group of patients with schizophrenia and healthy comparison subjects who were balanced by sex.

(14iii) This study examines whether subcortical volumes of the basal ganglia and thalamus in schizophrenic patients are related to neuroleptic exposure and symptom severity.

{ob1, ¬ ob4, ¬ ob8, …} (cf. 20vii)

(15i) The authors explored whether abnormal functional lateralization of temporal cortical language areas in schizophrenia was associated with a predisposition to auditory hallucinations and whether the auditory hallucinatory state would reduce the temporal cortical response to external speech.

(15ii) The authors designed a randomized, double-blind, crossover study to assess the efficacy of sertraline in the treatment of premenstrual dysphoric disorder (PMDD) when given only during the luteal phase of the menstrual cycle.

(15iii) We evaluated the antidepressant and mood-stabilizing effects of lamotrigine, a novel anticonvulsant, in a group of rapid-cycling bipolar patients (DSM-IV).

(15iv) We present various economic and statistical approaches to obtaining the required estimates.

{¬ ob1, ¬ ob4, ob8, …} (cf. 20v)

(16i) In this article, a speaker-based approach to aspect is proposed which crucially invokes abstraction, namely the idealization of different types of situations.

(16ii) In this paper, the simple context of the Wilcoxon-Mann-Whitney (WMW) test is used to illustrate alternatives where “one distribution is to the right of the other”.

(16iii) The aim of this paper is to characterize the hydraulic flow regime of an existing sedimentation tank and to simulate the hydraulic flow regime of an existing sedimentation tank.

(16iv) This paper presents the results obtained by "G.I.S. MOUSSES", which comprise the Cemagref and the six main private companies operating wastewater treatment plants (WWTP) throughout France.

(16v) This paper presents a discourse-functional account of English inversion, based on an examination of a large corpus of naturally-occurring tokens.

(16vi) This paper discusses the procedure used and the results obtained to optimize the combination of the storage volume and treatment rate to meet the required objectives.

{ob1, ¬ ob4, ob8, …} (cf. 20iii)

(17) In this paper, we describe a framework for developing probabilistic classifiers in natural language processing.

{ob1, ob4, ¬ ob8, …} (cf. 20iv)

(18) In this study we tested the hypothesis that size variation among larvae also serves as a cue triggering development of the cannibalistic phenotype.

{¬ ob1, ¬ ob4, ¬ ob8, …} (cf. 20viii)

(19i) The efficacy of extended release physostigmine salicylate, an acetylcholinesterase inhibitor, was evaluated in 850 subjects with mild-to-moderate Alzheimer disease (AD) in a multicenter trial.

(19ii) Treatment of an aquacultural wastewater from alligator farms in Louisiana using land application was investigated.

(19iii) Local land use, runoff and chemical data were used to estimate pollutant loads from non-point source pollution and pollutant loads by land use in three sub-basins of the Santa Monica Drainage Basin: namely, Ballona Creek, Malibu Creek, and Topanga Creek.

(19iv) An account of these phenomena is formulated in the context of COGNITIVE GRAMMAR.

Figure 4. Text-sentences filling the genre category [announcing the research].

Castel et al. (2000), Castel (2001, 2004, 2005b), and Hlavacka (2004) are studies which attempt to correlate field properties with lexicogramatical patterns in a principled way, i.e. to account for the regularities underlying the varied realizations of the genre category [research announcement] exemplified in Figure 4.

Castel (2001) specifically addresses the issue of formally mapping configurations based on decisions to refer or not to objects such as the *researcher* (**ob1**), the *research* (**ob4**), the *article* (**ob8**), etc. onto specific semantic processes and their associated participant roles. The configurations of objects heading the examples (14)-(19) are part of (20):

(20i) *{ob1, ob4, ob8, …}

(20ii) *{¬ ob1, ob4, ob8, …}

(20iii) {ob1, ¬ ob4, ob8, …}

(20iv) {ob1, ob4, ¬ ob8, …}

(20v) {¬ ob1, ¬ ob4, ob8, …}

(20vi) {¬ ob1, ob4, ¬ ob8, …}

(20vii) {ob1, ¬ ob4, ¬ ob8, …}

(20viii) {¬ ob1, ¬ ob4, ¬ ob8, …}

These configurations result from considering all the logically possible situations with and without (¬) explicit reference to the researcher (**ob1**), the research (**ob4**), and the article (**ob8**). The asterisk in (20i-ii) indicates that no single clause representatives of these configurations were found in the RedACTe Project's RAA sample.

Hlavacka (2004) is a more refined description of field features as they predetermine semantic features which in turn predetermine lexical and syntactic patterning. I now reconceptualize this description in the referentially oriented perspective of Castel (2001).

The *researcher* (**ob1**) defines the *research* (**ob4**) to answer a (scientific) question. In defining the research (**ob4**), the researcher (**ob1**) assigns the objectives, hypotheses and methodology to it (**ob4**) that he (**ob1**) believes are necessary in order to answer the question. Thus, the research (**ob4**) has objectives, hypotheses, and methodology which are the researcher's (**ob1**) objectives, hypotheses and methodology. Notice that th e objectives, hypotheses, and methodology are

statements about *domain-specific objects and relationships between objects*, and thus it is expected that reference to such objects and relationships between objects be made in the research announcement. Contrasted with the researcher (**ob1**), the research (**ob4**), and the article (**ob8**), these domain-specific objects and relationships between objects constitute an open set and explicit reference to them is obligatory.

Once the researcher's (**ob1**) objectives are attained (because the hypotheses are substantiated in accordance with the methodology), he (**ob1**) communicates the *results* (**ob7**) to potential readers. The results (**ob7**), i.e. the answer to the question initially posed, are also domain-specific objects and relationships between objects.

Since the researcher (**ob1**) defines the research (**ob4**) to answer a question, the research (**ob4**) is *logically* the investigation carried out by the researcher (**ob**1) to answer the question. Under certain circumstances, however, the researcher can reconceptualize the research (**ob4**) as if it were he himself (**ob1**).

Since the researcher (**ob1**) creates the *article* (**ob8**) to communicate the research (**ob4**), the article (**ob8**) is *logically* the location where the researcher (**ob1**) communicates (the results (**ob7**) of) the research (**ob4**). However, under certain circumstances, the researcher (**ob1**) can reconceptualize the article (**ob8)** as if it were he himself (**ob1**).

In the text location identified as [research announcement], the researcher (**ob1**) *basically* informs potential readers about the objectives and/or the hypotheses and/or the methodology that he (**ob1**) has assigned to the research **ob4**, i.e. **ob1** announces his definition of the domain-specific objects and relationships between objects which make up the research (**ob4**). Crucially, this announcement involves a reduced set of relationships between the above mentioned objects. I assume, following Fawcett (in preparation), that a series of LFs is required as input representations to the generation process itself. I assume, crucially, that these LF representations are defined for each terminal genre category by an appropriate set of rules. I now specifically illustrate this proposal in relation to the terminal genre category [research announcement].

Let the set of implications in Figure 5 represent a simplified version of the algorithm called by the terminal genre feature [research announcement] (cf. cc_lp4) to guide the user of the system in choosing an LF for this text location. The algorithm is used by the system to help the user decide an appropriate CC from the set of LFRs in Figure 6. LFRs are currently being developed by Luis

A. París and myself. I can give here only a very general idea of how these rules operate and interact with the rest of the system.

We need the LFRs in Figure 6 to define the appropriate LFs underlying the generation of text-sentences like (14)-(19) in the text location labelled [research announcement]. These text-sentences are instances of the lexicogrammatical potential associated with the terminal genre category [research announcement]. I share the CLG assumption that most of the semantic choices are predetermined by various decision algorithms operating on LF. I now show how the Research Announcement LF Algorithm controls LF Selection and, therefore, also LF Projection. The net effect of the decisions made on the basis of this algorithm is a predefined semantics which in turn predefines a syntactic representation which, after stripping, is the text proper.

Factors conditioning LF construction
annc_dalg1: annc_ornt_alg ⇒ impl [alg_tbd1] / expl.
annc_dalg2: impl ⇒ rch_ornt / art_ornt.
annc_dalg3: expl ⇒ rchr_as_such [alg_tbd1] / rchr_as_rch [annc_dalg_lp3] /
rchr_as_art [annc_dalg_lp4].
annc_dalg4: rhcr_as_such ⇒ (rch_as_such / art_as_such) &
(prp_dmtd [annc_dalg_lp1] / prp_pmtd [annc_dalg_lp2]).
... (Other implications.)

Auxiliary algorithms
alg_tbd1: (It defines which object ends up later as the subject-theme.)
... (Other auxiliary algorithms.)

LF predetermination
annc_dalg_lp1: prp_dmtd & (rch_as_such / rchr_as_rch) ⇒ for lgc_entity prefer cmplx, ob4,
pd = lg_act_pd, ob1= Ag, ob4 = Cre.
annc_dalg_lp2: prp_pmtd & (art_as_such / rchr_as_art) ⇒ for lgc_entity prefer smplx, ob8,
pd = lg_cg1_pd, ob1 = Ag-Cg, ob_n = Ph, ob8 = Lc.
annc_dalg_lp3: rchr_as_rch ⇒ for lgc_entity prefer smplx, ob4, pd = lg_cg1_pd, ob4 = Ag-Cg, ob_n = Ph.
annc_dalg_lp4: rchr_as_art ⇒ for lgc_entity prefer smplx, ob8, pd = lg_cg2_pd, ob8 = Ag-Cg, ob_n = Ph.
... (Other LF predetermination rules.)

Figure 5. Research Announcement LF Selection Algorithm.

LF Selection
lf1: lgc_entity ⇒ (object / event [lf_lp1]) & (smplx [lf_lp3] / cmplx [lf_lp4]).
… (LF Selection Rules for object and predicate insertion and role assignments.)

Inter-stratal LF Projections
lf_lp1: event ⇒ for entity prefer congruent_situation, independent, ...
lf_lp2: object ⇒ for entity prefer thing, ...
lf_lp3: smplx & lg_cg2_pd ⇒ for entity prefer causing_self_to_come_to_know_thing, ...
lf_lp4: cmplx & lg_act_pd ⇒ for entity prefer rlv_plus_created, proposition_cre, rlv_cog_dep, ...
… (Other LF Projection Rules.)

Key: rch = research; alg = algorithm; impl = implicit; expl = explicit; ornt = oriented; rchr = researcher; art = article; prp = purpose; dmtd = demoted; pmtd = promoted; dalg = decision algorithm; tbd = to be defined; cmplx = complex; smplx = simplex; lg_act_pd = variable ranging over the relevant class of action predicates corresponding to the relevant action processes in the CLG semantics for verbs such as *design*, *carry out*, etc.; Cre = Created; Ag-Cg = Agent-Cognizant; Ph = Phenomenon; lg_cg1_pd = variable ranging over the relevant subclass of cognition predicates corresponding to the relevant cognition processes in the CLG semantics for verbs such as *examine*, *test, explore*, etc.; lg_cg2_pd = variable ranging over the relevant subclass of cognition predicates corresponding to the relevant cognition processes in the CLG semantics for verbs such as *discuss*, *present, formulate*, etc,; rlv_plus_created = variable ranging over the relevant subclass of CLG action processes; proposition_cre = feature that makes the action process subcategorize a dependent situation; rlv_cog_dep = variable ranging over preferences for the appropriate cognitive process of the dependent clause.

Figure 6. LF Rules.

Having reached the point where the terminal genre category [research announcement] needs to be associated with an LF representation, the user is invited to interact with the system so that a CC can be assigned to this text location. Thus, the user is expected to choose, in accordance with the LFR lf1, between an 'object' and an 'event', and between 'simplex' and 'complex'. Since this choice is not made in a vacuum but within the terminal genre category [research announcement], the system automatically takes care of two fundamental aspects: (i) the application of the Inter-stratal Projection Rule cc_lp4, on the basis that it is this terminal category and not, say, [title], and (ii) the calling of the relevant algorithm, annc_dalg1-4, to assist him with the decisions involved in constructing an LF. The application of cc_lp4 enforces the selection of 'event'. Let us now see how the algorithm enforces other LF choices.

The user is asked whether he wants to write a research announcement by orienting it explicitly ('expl') or implicitly ('impl') to the research activity carried out or the article where the results are communicated.

If he chooses 'impl', then the relevant LFRs introducing the research (**ob4)** and the article (**ob8)** are pre-set so that these objects are not introduced into the LF representation; therefore, there will be no explicit reference to **ob4** or **ob8** in the text-sentences realizing such LF.[24] Now, the LFRs introducing the event predicate must be applied. The predicate type and subtype are chosen on the basis of whether the user accepts the algorithm's offer to implicitly orient the research announcement either to the research or to the article. If he then chooses the research-orientation

('rch_ornt'), a cognitive predicate is introduced, on the basis of other LFRs, with the researcher (**ob1**) as the Agent-Cognizant and a domain-specific object (**ob$_n$**) as the Phenomenon. Examples of exponents of this class of cognitive predicates are *assess*, *examine*, *explore*, *test*, etc. (henceforward, the *examine* subtype). Cf. text-sentences (15i, iii) and (19i-iii). Notice that the choice of 'impl' triggers the application of another algorithm, alg_tbd1, whose task is to help the user decide which of the two objects, **ob1** or **ob$_n$**, will be the subject-theme. However, if he chooses the article-orientation ('art_ornt'), a cognitive predicate of the appropriate subtype is inserted, on the basis of other LFRs, with the researcher (**ob1**) as Agent-Cognizant and a domain-specific object (**ob$_n$**) as the Phenomenon. Examples of exponents of this subtype of cognitive predicates are *discuss*, *present*, *formulate*, etc. (henceforward, the *discuss* subtype).[25] Cf. text-sentences (15iv), (19iv). The algorithm alg_tbd1 is also called to decide which of these two objects will be the subject-theme.

If the user decides to explicitly orient the announcement ('expl'), then the research (**ob4**) or the article (**ob8**) is inserted in the LF representation under construction. If 'rch_as_such' or 'rchr_as_rch' is chosen, then **ob4** is inserted. If 'art_as_such' or 'rchr_as_art' is chosen, then **ob8** is inserted.

The consequence of the path [expl, rchr_as_such, rch_as_such, prp_dmtd] is that the LFs underlying text-sentences like (14ii) and (15ii) are defined. The decision to choose to demote the announcement of the specific contribution ('prp_dmtd'), triggers the application of a projection rule (annc_dalg_lp1) whose effect is to make the LF under construction a complex event with an action predicate of the subtype corresponding to verbs such as *design*, *carry out*, etc. and to assign the role of Agent to the researcher (**ob1**) and the role of Created to the research (**ob4**). Furthermore, an event dependent on this action predicate is also defined so that the domain-specific object (**ob$_n$**) is assigned the role of Phenomenon of a subordinate cognitive predicate. Examples of the dependent cognitive predicates are, again, verbs of the *examine* subtype. As in the 'impl' cases, the algorithm alg_tbd1 is called to define whether the researcher (**ob1**) or the research (**ob4**) will end up as the subject-theme of the main event.

If the path chosen is [expl, rchr_as_such, rch_as_such, prp_pmtd], then the LFs of text-sentences like (14i) and (18) are defined. The feature 'prp_pmtd', which reflects the user's decision to highlight his specific contribution, triggers the application of the projection rule annc_dalg_lp2.

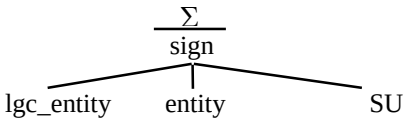
This rule defines the LF as a simplex event with a cognitive predicate whose Agent-Cognizant and Phenomenon are the researcher (**ob1**) and the domain-specific object (**ob$_n$**), respectively. Another effect of the algorithm is to associate the research (**ob4**) with the role of Location so that it eventually ends up as an appropriate Adjunct of the main clause. The auxiliary algorithm alg_tbd1 is called to decide whether **ob1** or **ob$_n$** will be the subject-theme of the main clause.

Similarly, the paths [expl, rchr_as_such, art_as_such, prp_dmtd] and [expl, rchr_as_such, art_as_such, prp_pmtd] account, respectively, for the LFs underlying text-sentences like (16ii) and (16i, 17). The subject-theme resolution algorithm also intervenes to decide which object ends up as the subject-theme. In both cases, the article (**ob8**) is assigned the role of Location.

The paths [expl & rchr_as_rch] and [expl & rchr_as_art] allow for the construction of the LFs corresponding to text-sentences like (14iii) and (16iii-iv, vi), respectively. They are both defined as simplex events by the projection rule annc_dalg_lp2. They differ in the object (**ob4** or **ob8**) that replaces the researcher (**ob1**) in the role of Agent-Cognizant of the cognitive predicate. If 'rchr_as_rch' is chosen, then the exponents of the class of cognitive predicates are verbs of the *examine* subtype. If 'rchr_as_art' is chosen, then the exponents end up being verbs of the *discuss* subtype. Notice that the object replacement enforced by the features 'rchr_as_rch' and 'rchr_as_art' makes the application of the auxiliary subject-theme resolution algorithm unnecessary.

Except for the very general and uninformative object variable **ob$_n$**, the above LFs do not specify the reference to domain-specific objects and relationships between objects. These more delicate referential specification should also be part of the LFs which are constructed for each terminal genre category. It is these fully referentially specified LFs headed by [lgc_entity] in (21) which predetermine semantic choices within [entity].

(21)



The Inter-stratal LF Projection Rules are predetermination rules in the sense that they make certain semantic feature configurations obligatory by altering the default probabilities associated

with the component features of SNRs. A rule like lf_lp4, for example, modifies the SNRs involved in defining (a few aspects of) the matrix selection expression of text-sentences like (14ii).

CONCLUSION

I have shown how to account for systematic correlations between the RAA GSP and LGP by re-sorting to a genre-controlled approach to FSP construction. This approach is formalized by (i) introducing into CLG the new unit (*linguistic*) *sign*, and (ii) extending CLG basic rule mechanism so that components higher that the lexicogrammatical component are also defined as sets of implications. Terminal genre categories are conceptualized as text locations where LF representations of relationships between objects are largely predetermined by decision algorithms defined on the basis of genre-driven properties of various sorts. Depending on which terminal genre category needs to be filled with a linguistic sign, the Inter-stratal Genre Projection Rules call the relevant algorithms for deciding LF choices. The resulting LF representations are mapped, by the Inter-stratal LF Projection Rules, onto semantic representations which in turn are mapped onto syntactic representations which in turn are mapped onto text proper.

The approach was illustrated with descriptive results of the RedACTe Project, particularly, with generalizations about the LGP associated with the terminal genre category [research announcement] as proposed in Hlavacka (2004). There are two main subtypes of cognitive processes involved in the realization of this genre category: the *examine* subtype and the *discuss* subtype (cf. §3.2). The choice of one or the other depends on the research announcement orientation that the researcher-writer wishes to adopt in writing this RAA text location. These two subtypes, furthermore, interact in interesting ways with decisions related to whether to make explicit or not the orientation adopted. Thus, if an explicit research-oriented announcement is chosen, the writer is algorithmically assisted to choose either hiding himself behind the research (in which case the research will end up as the Agent of a cognitive process of the *examine* subtype), or competing with the research in the participant role assignments (in which case the researcher will end up either as the Agent of an action process of the *design* subtype of the main clause if the decision is made to demote the specific contribution, or as the Agent of a cognitive process of the *examine* subtype if the decision is made to promote the specific contribution).

While corpus-based, and thus descriptively interesting, and formally defined, and thus testable, the generalizations underlying these results are valid only for one terminal genre category of the RAA, namely the Reseach Announcement. Furthermore, the generalizations are neither complete nor delicate enough to allow for the development of a sound generation system. Both limitations, however, can be overcome with extensive descriptive work on the systematic correlations between culture and language which account for the textual realization of the other terminal genre categories of the RAA GSP (cf. Figure 2). It is the belief of the RedACTe Project that only after a "complete and delicate" generation-oriented description of all the RAA terminal genre categories is provided, will its long-term goal be attained of developing a robust computational tool capable of assisting scientific researchers in the writing of RAAs.

Acknowledgements

The research underlying the results presented here was supported by Grant #06/G256 from the Secretaría de Ciencia y Técnica de la Universidad Nacional de Cuyo, Mendoza, Argentina. A version of the main ideas of the paper was communicated at a plenary session of the Second Latin American Systemic Functional Linguistics Conference, Concepción, Chile, 17-18 November, 2005. I am grateful to Ana M. Miret for valuable comments on various aspects of the paper.

REFERENCES

Bazerman, Ch. (1988) *Shaping Written Knowledge. The Genre and Activity of the Experimental Article in Science*. Wisconsin: The University of Wisconsin Press.

Boccia, C. (2001) *Propuesta de análisis retórico-lingüístico de la Introducción del artículo de investigación científica en inglés desde una perspectiva Sistémico-Funcional*. Tesis no publicada para optar al título de Magíster en Lingüística Aplicada, Universidad Nacional de Cuyo, Mendoza, Argentina.

Car, E. C. (2000) *La estructura del abstract del artículo de investigación científica en inglés: una adaptación del modelo Create a Research Space de Swales*. Tesis no publicada para optar al título de Magíster en Lingüística Aplicada, Universidad Nacional de Cuyo, Mendoza, Argentina.

Castel, Víctor M. (2006a) *A Computer Implementation of the Cardiff Grammar Generator*. Computational Linguistics Unit, University of Cardiff, Wales, United Kingdom. For access permission contact rp.fawcett@virgin.net.

Castel, Víctor M. (2006b) *The Cardiff Grammar Generator Online Help*. Computational Linguistics Unit, University of Cardiff, Wales, United Kingdom, and INCIHUSA, CONICET, Mendoza, Argentina. For access permission contact vcastel@lab.cricyt.edu.ar.

Castel, V. M. (2005a) "Determinación de valores de verdad de condiciones de reglas de generación de textos". In V. M. Castel (Comp.) *Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos* (pp. 23-34). Mendoza: Editorial de la Facultad de Filosofía y Letras de la Universidad Nacional de Cuyo. Soporte Internet.

Castel, V. M. (2005b) "Generación de artículos de investigación científica: el *abstract* como una estructura de rasgos sistémicos funcionales". *Signo & Seña*, 14, 257-282.

Castel, V. M. (2004) Towards a Generation-oriented Grammar of the Research Paper Abstract. In S. Hassan (Ed.) In *Actas* del Primer encuentro latinoamericano de Lingüística Sistémica Funcional: La lengua y la educación (pp. 44-57). Mendoza: Editorial de la Facultad de Filosofía y Letras, UNCuyo. Soporte CD.

Castel, V. M. (2001) Proyección de configuraciones retóricas en configuraciones de transitividad en el *Abstract* del artículo de investigación científica en inglés. In H. Albano, L. Ferrari & M. Giammatteo (Coord.) *La Gramática: Modelos, Enseñanza, Historia* (pp. 181-210). Buenos Aires: Instituto de Lingüística de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires.

Castel, V. M. & Diblasi, A. (1999) Distribución disciplinar de los participantes del *Abstract* del artículo de investigación científica en inglés. *Revista Argentina de Lingüística*, 15(1), 83-105.

Fawcett, R. P. (2000) *A theory of syntax for systemic functional linguistics*. Amsterdam: John Benjamins.

Fawcett, R. P. (In preparation) *Alternative Systemic Functional Architectures: How do we choose?* London: Equinox.

Fawcett, R. P., Tucker, G. H. & Lin, Y. Q. (1993) How a systemic functional grammar works: The role of realization in realization. In H. Horacek & M. Zock (Eds.) *New concepts in natural language generation* (pp. 114-86). London: Pinter.

Halliday, M. A. K. (1987) Language and the Order of Nature. In Fabb et al. (1987: 135-154). N. Fabb, D. Attridge, A. Durant and C. MacCabe. Eds. *The Linguistics of Writing: Arguments between Language and Literature* (pp. 135-154). Manchaster: Manchester University Press.

Halliday, M. A. K. & Matthiessen, C. M. (1999) *Construing Experience Through Meaning*. London: Continuum.

Hasan, R. (1984, 1996) The Nursery Tale as Genre. C. Cloran, D. Butt & W. Geoffrey, Eds. *Ways of Saying: Ways of Meaning. Selected papers of Ruqayia Hasan* (pp. 51-72). London: Cassell.

Hlavacka, L. E. (2004) *Patrones de transitividad en la sección Anuncio de la Investigación del Abstract del artículo de investigación científica en inglés: descripción sistémico-funcional y sistematización de los valores de campo que determinan la selección de procesos verbales y configuraciones asociadas*. Tesis no publicada para optar al título de Magíster en Lingüística Aplicada, Universidad Nacional de Cuyo, Mendoza, Argentina.

Mann, W., Matthiessen, C. M. & Thompson, S. A. (1992) Rhetorical structure theory and text analysis. In W. Mann & S. A. Thompson (Eds.) *Discourse Description: diverse analyses of a fund raising text* (pp. 39-78). Amsterdam: John Benjamins.

Martin, J. R. (1992) *English Text - System and Structure*. Amsterdam: John Benjamins.

Matthiessen, C. M. (1993) Register in the round: diversity in a unified theory of register analysis. In M. Ghadessy (Ed.) *Register Analysis. Theory and Practice* (pp. 221-292). London: Pinter.

Miret, A. M. (2002) *Estructura genérica de la sección Discussion en artículos de investigación científica en Medicina*. Tesis no publicada para optar al título de Magíster en Lingüística Aplicada, Universidad Nacional de Cuyo, Mendoza, Argentina.

Miret, A. M. (In preparation) *Hacia un modelo de realización lexicogramatical de la Conclusión en el Abstract del artículo de investigación en Medicina*. Tesis para optar al título de Doctorado en Letras, Universidad Nacional de Cuyo, Mendoza, Argentina.

París, Luis A. y V. M. Castel (2005) "La relación retórica Solución en el *Abstract* del artículo de investigación científica". In *Actas* del VI Congreso Latinoamericano de Análisis del Discurso, Pontificia Universidad Católica de Chile, Santiago, Chile, 2005. Available at http://www.congresoaled2005.puc.cl/fset_actas.html.

Swales, J. (1990) *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Weissberg, R. & Buker, S. (1990) *Writing Up Research. Experimental Research Report Writing for Students of English*. Englewood Cliffs, N.J.: Prentice Hall Regents.

Rezzano, S. (1999) *Modalidad de probabilidad y evidencia en el artículo de investigación científica en inglés*. Tesis no publicada para optar al título de Magíster en Lingüística Aplicada, Universidad Nacional de Cuyo, Mendoza, Argentina.

Rezzano, S. (2003) Modality and Modal Responsibility in Research Articles in English. In R. Facchinetti & F. Palmer (Eds.) *English Modality in Perspective: genre analysis and contrastive studies*. Bern: Peter Lang.

Zeevat, H., Klein, E. & Calder, J. (1987) Unification Categorial Grammar. J. Haddock, E. Klein & G. Morill (Eds.) *Categorial Grammar, Unification Grammar and Parsing* (pp. 195-222). Edinburgh: Centre for Cognitive Science.

NOTES

[1] For a computational implementation of CLG, see Castel (2006a, b).
[2]On the use of the term 'sign' here, I have been influenced by both Saussure's original concept, and Zeevat, Klein & Calder (1987)'s interpretation of it in the framework of Unification Categorial Grammar.
[3]'qlg' = quality group and 'qtg' = quantity group are the other CLG units.
[4]A disjunct may itself be a conjunction, and a conjunct may itself be a disjunction, so that condition *p* can indeed be fairly complex. Cf. (5iv-v).
[5]I am deliberately ignoring another subtype of RR, namely, Adjustment Rules, for they play no role in the main issue discussed in this paper.
[6]The introduction of a feature is the result of a choice made by the user of the system or the system itself depending on whether the system is functioning under interactive or random generation, respectively.
[7]According to Fawcett (personal communication), SNRs and SPRs can be viewed as explicit and implicit system network rules, respectively.
[8]Or, more precisely, the equivalent of this rule in RedACTe.
[9]For the algorithms involved in determining the truth value of condition *p*, see Castel (2005a).
[10]For an interesting presentation and review of various types of SFL architecture, see Fawcett (in preparation).
[11]A 'terminal' genre category is any genre feature whose daughters are not genre features.
[12]'U' followed by a number is used to facilitate the reference to the different units which make up this RAA.

[13]See Fawcett (in preparation) for an insightful discussion about the need to define higher level algorithms for deciding how to choose system network features. See also the algorithm in Figure 5.
[14]Under interactive generation, RedACTe makes a distinction between two sets of feature: underlying features, for use by the generator, and interface features, for a friendly presentation to the user of the system. Of course, all features in rules in Figures 2-3 are underlying features.
[15]This is our formal way of saying that RedACTe is incapable of accounting for other text types.
[16]I use [title] in the text instead of [wth_ttl], etc. to facilitate readability.
[17]Essential aspects of these generalizations are taken from Miret (in preparation).
[18]But see representation (21) below which is actually the type of structure defined by cc_lp1, where there is a place for the unit 'lgc_entity'.
[19]This definition is more complex and controversial than what I am making it appear.
[20]This KB appears to correspond to what in Halliday & Matthiessen (1999) are three knowledge bases: the ideational, the interpersonal and the textual bases, and to what Fawcett (in preparation) calls the belief system.
[21]I assume that the referential function of language roughly corresponds to what in SFL is known as the ideational metafunction, so including both the experiential and the logical subfunctions. The avoidance of the term 'reference' by leading systemic functional linguists (Halliday, 1987; Martin, 1992; Matthiessen, 1993) in the context of the ideational metafunction seems to reflect their preference for the "rhetorical-ethnographic" orientation over the "logico-philosophical" orientation (cf. Halliday & Matthiessen, 1999: 415-6).
[22]Only units U0-U6 are analysed here. For a slightly more complete analysis of (7), see París & Castel (2005).
[23]These readers are typically other researchers of the research community to which **ob1** belongs, especially, researchers investigating issues related to the CC introduced by **ob0**, among which the researchers **ob3** occupy the position of highly qualified experts.
[24]Though there is neither reference to the research nor the article, these cases are classified as research-oriented and article-oriented because the intervening predicates are of the same subtypes as the predicates underlying text-sentences which do explicitly refer to the research or the article, respectively. See below.
[25]Hlavacka (2004) classifies these verbs as **verbal** processes.