Mechanisms of instantiation: towards an account of how language becomes text

Giacomo Figueredo, Francieli Silvéria Oliveira, Eduardo José da Silva Luz

https://doi.org/10.1590/SciELOPreprints.16753

Submitted on: 2026-07-01
Posted on: 2026-07-01 (version 1)
(YYYY-MM-DD)

Mechanisms of instantiation: towards an account of how language becomes text

Mecanismos de instanciação: uma descrição acerca de como a língua se torna texto

Giacomo Figueredo
Universidade Federal de Ouro Preto
giacomo.figueredo@ufop.edu.br
https://orcid.org/0000-0002-0967-244X
Financiamento CNPq Nº do processo: 310547/2023-7

Francieli Silvéria Oliveira
Universidade Federal da Paraíba
francieli.oliveira@academico.ufpb.br
https://orcid.org/0000-0002-4895-1919

Eduardo José da Silva Luz
Universidade Federal de Ouro Preto
eduluz@ufop.edu.br
https://orcid.org/0000-0001-5249-1559
Financiamento CNPq Nº do processo: 308400/2022-4

Resumo: Este artigo parte de uma distinção central na Linguística Sistêmico-Funcional, a relação entre as dimensões da instanciação e da estratificação, para investigar como o 'texto como instância' constitui um domínio descritivo próprio, correlacionado ao 'texto como língua em contexto', embora sendo desse distinto. Nessa perspectiva, concebe-se o texto como uma configuração probabilística de escolhas sistêmicas, cuja variação entre registros não se explica apenas pelo potencial do sistema, mas por mecanismos que operam o comportamento do sistema, cujo domínio é a instanciação. Para investigar tais mecanismos, compilou-se um *corpus* em português brasileiro, organizado a partir de três famílias de gêneros: avaliar (artigos acadêmicos), informar (cartilhas educativas) e envolver (narrativas orais). A análise identificou três mecanismos de instanciação. A redistribuição descreve o deslocamento das probabilidades de seleção de um tipo de texto para outro. A chavicidade associativa localiza redes de cosseleção próprias de cada tipo de texto. A dinâmica de sistemas modela como a instanciação se desenvolve ao longo do tempo do texto, revelando como a restrição sistêmica não é uma medida estática. Os resultados indicam que a tipologia de textos decorre de padrões localizados de cosseleção sistêmica, e não apenas da frequência de categorias isoladas. O estudo sustenta que instanciação e estratificação devem ser mantidas como dimensões complementares, porém não redutíveis entre si.
Palavras-chave: Instanciação; Gramática Probabilística; Logogênese; Comportamento Linguístico; Tipologia de Registro.

Abstract: The paper interrogates the theoretical distinction between instantiation and stratification in Systemic Functional Linguistics by treating 'text as instance' as a probabilistic configuration – one that is related to, yet distinct from, 'text as language in context'. While

standard SFL accounts view instantiation primarily as the manifestation of system potential, we complement this view by posing that instantiation constitutes a discrete domain of constraint in which systemic resources are probabilistically redistributed. In order to do so a specialised corpus in Brazilian Portuguese was compiled across three genre families: Evaluating (academic research articles), Informing (educational booklets), and Engaging (oral narratives). The analysis reveals three instantiation mechanisms: redistribution, whereby register constraints reframe baseline system probabilities; associative keyness, which locates co-selection networks; and system dynamics, showing that instantiation unfolds as a conditional probabilistic patterning rather than a static index. The findings show that text typology is driven by localised patterns of systemic co-selection rather than mere frequency, and offer a replicable methodology for tracking how systemic constraint scales up into text types. The paper argues that instantiation and stratification must be maintained as complementary but non-reducible dimensions.
Keywords: Instantiation; Probabilistic Grammar; Logogenesis; Language Behaviour; Register Typology.

Introduction

In Systemic Functional Linguistics (SFL), 'text' is conceptualised through its systemic organisation: linguistic phenomena are driven primarily by use (Halliday, 1978; Martin, 1992; Matthiessen, 2015), and meaning-making is inherently socially oriented (Painter, 1999). Because the system is probabilistic (Halliday, 1991, 1992), meaning-making has, inevitably, a probabilistic character. 'Text' is therefore a unit of discourse semantics – with communicative and social purpose – whose behaviour is an emergent property of the probabilistic relationship between language and context (cf. Matthiessen, 1993; Martin & Rose, 2008).

Complementarily, 'text' is also a fundamental unit of probability: a configuration of choices that associate probabilistically in the form of an instance. This complementarity extends to all strata – to the relationship between context and the complexity of activities and social relations (cf. Maton, 2014; Hao, 2020; Doran, Martin & Herrington, 2024; Doran, in prep.), and within language, between discourse semantics and grammar (cf. Nesbitt & Plum, 1988; Halliday, 1992; Matthiessen, 2015a; Figueredo et al., 2022; Hao & Wang, 2022).

Texts are distributed into distinct types, yielding a typology bounded by context. From the stratificational vantage point, texts are tied to context (register and genre): registers give texts a probable configuration, genres endow them with social purpose. This variation can also be viewed as a sub-potential (cf. Martin, 2010; Matthiessen, 2015a, 2015b): a set of probable resources available for selection from the system potential. From the instantiational vantage point, by contrast, text is observed as behaviour: rather than a sub-potential of resources, it is a configuration of probable choices organised by mechanisms of frequency distribution unfolding

in time. This view does not stand in opposition to the former; rather, it complements it by supplying patterned linguistic configurations for recurrent configurations of use (cf. Doran, Martin & Herrington, 2024; Doran, in prep.).

A one particular 'text' can be defined, relative to register and genre, as a linguistic unit that realises a set of variations in register, configured by a genre's social purpose (Martin, 1992; Martin & Rose, 2008; Martin, 2015), differing from other 'texts' – other sets of variations carrying on distinct social purposes (Plum & Cowling, 1987; Nesbitt & Plum, 1988; Zappavigna & Martin, 2018; Rose, 2019). At the same time, from the standpoint of probability, text typology is not given by sub-potentialisation alone: what matters is how systems co-occur and predict one another across instances – both probable and recurrent (Halliday, 1992) – since all texts are drawn from the same systemic resources, differing in the frequencies and proportions with which those resources are taken up (Firth, 1957; Ellis & Ure, 1969; Halliday, 1991; Martin & Christie, 1997). Certain choices weigh more heavily than others in setting text types apart (cf. Lemke, 1995; Rose, 2019).

The paper investigates instantiation and stratification as discrete domains, with 'text as instance' correlated with – but independent from – 'text as language in context'. The driving question is: from a strictly observational, probabilistic standpoint, what instantiational mechanisms does sub-potentialisation map onto – especially in its bearing on registers and genres (cf. Martin, 2009)?

To address this question, the paper does three things: offers a quantitative description of text as a unit of variation in systemic choices; sets out the mechanisms whereby choices settle into predictable patterns; and shows how this description feeds the view of text as a unit of probability thereby explaining why it divides into types. Such descriptions matter because the resources from stratification are identical across all text types – what sets types apart is (i) the uneven distribution of those resources, (ii) the weight each acquires, and (iii) the point at which it acquires that weight in the text's unfolding (Plum & Cowling, 1987; Halliday & James, 1993; Neumann, 2013; Oliveira, 2022; Figueredo et al., 2024).

Three hypotheses guide the analysis: (a) a quantitative approach is required to observe text behaviour in instantiation; (b) textual typology is an effect of the redistribution and association that resources assume under a given register configuration, expected to shift across genre families; and (c) this process runs dynamically, in temporal unfolding, so that redistribution and association are properties of instantiation rather than of the systems, which is what allows the instantiating of texts to be tracked as behaviour.

The scope is bounded by the sample. Claims about 'the language as a whole' are not made; the findings hold for the corpora described. As a result, throughout the paper, we reserve the technical terms 'sample' for texts-as-instances (the clusters of choice bound by conditioned probability) and 'corpus' for texts-as-resources (the units of discourse semantics functioning in context).

The corpus was compiled from a project on Self-Care in Diabetes Mellitus (DM), drawn from three sources that make up Sub-Corpus 1, SC1, Evaluating: academic articles; Sub-Corpus 2, SC2, Informing: educational booklets; and Sub-Corpus 3, SC3, Engaging: patient statements. Field is held constant across sub-corpora; genre family varies. Prior work mapped the broader domain across nine genres in three genre families (Oliveira, 2022, forthcoming; cf. Rose & Martin, 2012); this study takes up the three sub-corpora for instantiational analysis. The selection is governed by stratification – a random draw is the only unbiased one, so the corpora are stratificationally controlled – and read instantiationally as samples of sub-populations. The methodology is developed in Section 3 below.

2. Text in a Systemic Functional Framework

In SFL, the meanings in a text entail the simultaneous realisation of choices across every system of language – phonology, grammar, and discourse semantics alike (cf. Halliday & Hasan, 1976; Martin, 1992; Halliday, 1985; Halliday & Greaves, 2008). No system means on its own: a choice in one system hangs on selections made in others (Lemke, 1991; Halliday, 2005). The systems of each stratum realise their features through functional structures, and once unified within the unit 'text', these structures build, so to speak, a 'macrostructure' – a text structure made up of function structures realising all selected features of all language systems. Example (1) below, a clause from Sub-corpus 1 (EVA_05), shows a few systems operating simultaneously across strata; Figure 1 sets out its selected systems across the discourse-semantic and grammatical strata.

(1) [EVA_02, clause 16]:

A assistência de enfermagem	*objetiva*	*alcançar a melhoria*	*da qualidade de vida dos indivíduos.*
the nursing care	aims.3SG	to.achieve the improvement	of.the quality of life of.the individuals

'Nursing care aims to improve the quality of life of these individuals.'

Figure 1 – Example of selected systems of clause (1) across 19 systems

Stratum	**System**	**Selected feature**
discourse semantics	CONNEXION	no connexion
	LOGICO-SEMANTIC RELATION	expansion: extension
	TAXIS: parataxis	paratactic: primary
	TAXIS: hypotaxis	hypotactic: alpha
	MESSAGE: periodicity	no periodicity marking
	ENGAGEMENT	monoglossia
grammar	MOOD	indicative: declarative
	MODALITY	no modality
	PROCESS TYPE	material: transformative
	DEIXIS	present
	POLARITY	positive
	CIRCUMSTANCE	no circumstance
	THEME: textual	relative theme
	THEME: interpersonal	no interpersonal theme
	THEME: ideational	no ideational theme
	SUBJECT: person	non-interactant (3rd)
	SUBJECT: responsibility	responsible
	SUBJECT: number	singular
	SUBJECT: presumption	recoverable: explicit

Note: prepared by the authors.

In this sense, the structure of a text is not a matter of merely piling discourse semantic structures on top of grammatical structures on top of phonological ones. Rather, the semantic unification of the text binds structures together through co-selection among systems at all strata (cf. Martin, 2015). Co-selection, in these terms, runs in two directions at once. Vertically, it ties systems across strata: a choice in NEGOTIATION co-selects (i.e., expects and is expected by) a choice in SPEECH FUNCTION, which in turn co-selects (expects and is expected by) a choice in MOOD, and so on. Horizontally, it ties systems within a stratum – a choice in MOOD co-selects with choices in TRANSITIVITY and THEME – and within a rank: within the nominal group, choices in DEIXIS, EPITHESIS, CLASSIFICATION, and QUALIFICATION co-select. What finally enables systems to co-select is the social purpose of the text (Martin, 2015; Rose, 2019): coherence is the product of systemic organisation steered by purpose (Halliday & Hasan, 1976; Martin, 1992; Zappavigna & Martin, 2018).

This stratificational picture, however, leaves one central question unanswered. Stratification says how strata relate through realisation, but not how this abstract ‘text macrostructure’ comes to be actualised as a single, existing text – that is, how “all strata instantiate” (Martin, 2009, p. 54). It does not name the mechanism by which one configuration ends up being the one produced, rather than any of the others the system allows. That is the

question instantiation has to answer. Instantiation is precisely the principle that, from among the configurations the system permits, determines which one is realised in a given instance; and the way that choices in the system are patterned across instances is what we call 'language behaviour'.

Context, accordingly, must be grasped not simply as a stratum but as a set of systems that themselves enter into co-selection. Just as NEGOTIATION co-selects with discourse functions and MOOD, contextual systems – such as STATUS, CONTACT, and DEGREE OF EXPERTISE – co-select with them (cf. Oliveira, 2022), fixing the probabilistic, non-random principle behind systemic choices.

2.1 Instantiation and Quantitative Analysis

Where strata are linked by realisation, instances are linked by similarity in behaviour – the probabilities with which a system's features are chosen across instances. Types of text do not differ in the resources available to them; what sets types apart is how the shared resource behaves. Instances fall into groups according to how generalised their configurations are. The property enabling realisation is meta-redundancy (Lemke, 1989), or the encoding of meaning-making at rising levels of abstraction: {genre realised by {register realised by [discourse semantics realised by grammar (realised by phonology)]}}. Behaviour, for its part, is enabled by para-redundancy (Figueredo et al., 2024), or the limit of a probability region within which behaviour stays similar: {{[(this text) is research article] is exposition} is informative} (Figure 2).

Two properties of para-redundancy matter especially. First, systems do not switch their choices wholesale between text types; many hold onto similar probabilities across types – the probability of selecting [interrogative] MOOD in an educational booklet and in a research article, for instance, is virtually the same. Second, variation is not always linear: text types richer in [imperative] do not necessarily line up against those richer in [indicative], since choices in other systems always bear on how types are configured. A type of text is therefore a reweighting of selection probabilities: some hold steady across types (para-redundancy), others shift, and it is the pattern of shifts that gives a type its profile.

These properties motivate the hypothesis that certain choices weigh more than others: they distribute resources and associate them in ways that locate the text within a recognisable

type. Such recognition shows up in probability that reaches beyond a single system – in the correlation of systems (cf. Halliday, 1991; Lemke, 1995; Neumann, 2012; Martin, 2015). For systems and features of this kind we use the terms ‘key system’ and ‘key feature’, by analogy with ‘keywords’ in Corpus Linguistics (Scott, 1997; Scott & Tribble, 2006). Keyness is always tied to other systems through correlation – whence the term associative keyness (Figueredo et al., 2024).

3. Methodology

3.1 Corpus Compilation

The corpus consists of texts in Brazilian Portuguese from the domain of Self-Care in Diabetes Mellitus (DM), compiled to allow comparison across genre families. It is drawn from three sources, which make up three sub-corpora: SC1 Sub-corpus 1 – academic articles (Evaluating); SC2 Sub-corpus 2 – educational booklets (Informing); and SC3 Sub-corpus 3 – patient statements (Engaging). Mapping the broader domain and identifying the genres at work in it was carried out in prior work in the project (Oliveira & Figueredo, 2020; Oliveira, 2022); the present study takes up the three sub-corpora for clause-level analysis. The sub-corpora were compiled by members of the Laboratório Experimental de Tradução (LETRA) at the Faculdade de Letras, Universidade Federal de Minas Gerais.

Each sub-corpus contributes five texts, fifteen in all. The unifying criterion in each case is genre family – not a single type – so each sub-corpus is internally varied at the level of specific genre while remaining controlled at the level of genre family. Sub-corpus 1 draws on research articles in Health Sciences, retrieved from SciELO and Brazilian institutional repositories (USP, UFMG); since each is a macrogenre (Rose, 2014), only its Introduction was retained, as it carries the text’s main staging and is comparable in length to the other sub-corpus texts. Sub-corpus 2 draws on booklets published by Brazilian diabetes organisations and health portals; all instruct the reader on managing self-care, placing them within the Informing genre family. Sub-corpus 3 draws on patient statements, testimonies, and one interview published on diabetes.org.br and YouTube; the consistent feature is that all are produced by speakers living with diabetes for an audience in the same condition, placing them within the Engaging genre family (Table 1).

Table 1 – Research corpus: texts, sources, and sizes

Text	Title (Portuguese / English)	Source type	Tokens	Clauses
Sub-corpus 1 – Academic articles (Evaluating)			**2,472**	**159**
EVA_01	Diferenças entre mulheres e homens diabéticos no autocuidado com os pés e estilo de vida / Differences Between Women and Men With Diabetes in Foot Self-Care and Lifestyle	research article	458	34
EVA_02	Qualidade de vida, conhecimento e atitude após programa educativo para Diabetes / Quality of Life, Knowledge, and Attitudes Following an Educational Program for Diabetes	research article	636	43
EVA_03	Prevalência de diabetes mellitus e fatores associados em mulheres indígenas do Município de Dourados, Mato Grosso do Sul, Brasil / Prevalence of Diabetes Mellitus and Associated Factors Among Indigenous Women in the Municipality of Dourados, Mato Grosso do Sul, Brazil	research article	614	33
EVA_04	Adesão das pessoas com diabetes mellitus ao autocuidado com os pés / Adherence to Foot Self-Care Among People With Diabetes Mellitus	research article	365	25
EVA_05	Atividades de autocuidado e suas relações com controle metabólico e clínico das pessoas com Diabetes Mellitus 1 / Self-Care Activities and Their Relationship with Metabolic and Clinical Control Among People with Diabetes Mellitus	research article	399	24
Sub-corpus 2 – Educational booklets (Informing)			**2,822**	**252**
DIR_01	Dia Mundial do Diabetes / World Diabetes Day	booklet	448	41
DIR_02	Cuidados com os pés / Diabetic Foot Care	booklet	700	69
DIR_03	Prevenindo o diabetes: esclarecendo dúvidas sobre Doenças Crônicas não Transmissíveis / Preventing Diabetes: Answering Common Questions About Noncommunicable Diseases	booklet	575	47
DIR_04	Dicas e cuidados no controle e prevenção do diabetes / Tips for Diabetes Prevention and Management	booklet	538	48
DIR_05	Orientações nutricionais para diabéticos / Nutritional Guidelines for People With Diabetes	booklet	561	47
Sub-corpus 3 – Patient statements (Engaging)			**2,117**	**289**
ENG_01	Relato da viagem de motocicleta / A Motorcycle Trip: A Personal Account	patient statement	284	50
ENG_02	'O diabetes é um touro domável, não é um bicho de 7 cabeças' José Loreto, ator, 33 anos e diabético tipo 1 desde os 14 / "Diabetes Can Be Managed—It's Not as Scary as It Seems": José Loreto, Actor, Age 33, Living With Type 1 Diabetes Since Age 14	patient statement	657	87
ENG_03	Uma história de otimismo na convivência com o diabetes / A Story of Optimism in Living With Diabetes	patient statement	405	60
ENG_04	Convivendo bem com o DM1 / Living Well With Type 1 Diabetes	patient statement	404	53
ENG_05	O que seria da nossa vida se não fossem os desafios? / What Would Life Be Without Challenges?	patient statement	367	39
Total – 15 texts, 3 genre families			**7,411**	**700**

Note: prepared by the authors, from the corpus compiled at the Laboratório Experimental de Tradução (LETRA), Universidade Federal de Minas Gerais.

In the analysis that follows, the corpus is treated as a sample, in the sense standard to quantitative and Corpus Linguistics (Biber, 1993; McEnery & Hardie, 2012). The corpus samples the domain of self-care in DM; each sub-corpus samples a sub-population of it – the clauses realised within one genre family (Evaluating, Informing, or Engaging). Seen from

stratification, these three groupings are the sub-potentials introduced in Section 1; seen from instantiation, they are samples of sub-populations. The two descriptions run in parallel and are kept distinct, in line with the treatment of instantiation and stratification as discrete domains.

The analysis is descriptive rather than inferential (Gries, 2009): it characterises the behaviour of these samples and compares them with one another, rather than estimating the parameters of the population they are drawn from. No inferential statistics are applied – not because the corpus exhausts the language, but because the questions pursued here – how resources are redistributed, associated, and unfolded across the sub-corpora – are answered by describing and comparing the samples themselves.

3.2 Corpus Annotation and Pattern Identification

The first methodological step was to build an annotation table in spreadsheet software, with a dedicated column for each of the 19 systems under analysis, filled in through a data-validation device kept uniform across all texts. Systems annotated span grammar and discourse semantics: MOOD, POLARITY, DEIXIS, PROCESS TYPE, MODALITY, THEME (textual, interpersonal, and ideational), CIRCUMSTANCE, CONNECTION, LOGICO-SEMANTIC RELATION, TAXIS (parataxis and hypotaxis), SUBJECT (person, responsibility, number, and presumption), MESSAGE (periodicity), and ENGAGEMENT. The annotation follows the description of Brazilian Portuguese on its own right (Figueredo 2007, 2011, 2020, 2021a, 2021b, 2026; Figueredo et al. 2014; Figueredo & Freitas, 2026).

The choices for each clause were then recorded as a single profile – the set of features selected across all 19 systems – which served as the unit of analysis for the steps that follow. Table 2 shows this profile for one clause from each sub-corpus, so that the three registers can be compared feature by feature.

Table 2 – Systemic profile of one clause from each sub-corpus

System	**EVA_02 (Evaluating)**	**DIR_04 (Informing)**	**ENG_05 (Engaging)**
MOOD	declarative	jussive imperative	declarative
POLARITY	positive	positive	positive
TEMPORAL DEIXIS	present	not applicable	past
PROCESS TYPE	material: transformative	material: transformative	material: transformative
MODALITY	no modality	obligation	no modality

THEME: textual	relative theme	no textual theme	no textual theme
THEME: interpersonal	no interpersonal theme	no interpersonal theme	no interpersonal theme
THEME: ideational	no ideational theme	no ideational theme	no ideational theme
CIRCUMSTANCE	no circumstance	purpose	temporal
CONNECTION	no connection	no connection	additive
LOGICO-SEMANTIC	extension	no relation	extension
TAXIS	hypotactic: alpha	no clause complex	paratactic: primary
SUBJECT: person	non-interactant (3rd)	addressee (2nd)	speaker (1st)
SUBJECT: responsibility	responsible	responsible	responsible
SUBJECT: number	singular	singular	singular
SUBJECT: presumption	recoverable: explicit	recoverable: implicit	recoverable: implicit
ENGAGEMENT	monoglossia	monoglossia	monoglossia

Note: prepared by the authors.
The example clauses are:
EVA_02 'Nursing care aims to improve the quality of life of these individuals';
DIR_04 'Diversify your diet for the good functioning of the organism';
ENG_05 'We set off at half past four in the morning'.

Reading Table 2 column by column gives the profile of each clause; reading it row by row shows where the registers agree and where they diverge. Several systems agree across all three (POLARITY positive, the THEME systems, ENGAGEMENT monoglossic); the sharpest differences fall in MOOD, MODALITY, and SUBJECT person. These profiles were then loaded into AntConc and searched as recurrent patterns, using the software as a frequency counter applied to systemic features rather than words. Pattern frequencies were tallied manually per text. This step identified which multi-system configurations recur across texts, and at which points in each text's unfolding – the basis for both the co-selection analysis (Section 3.3) and the dynamic model (Section 3.4).

3.3 Co-selection Analysis

Each clause simultaneously selects options across all 19 systems. Beyond describing which options are most frequent (the profile established in Section 3.2), the analysis asked whether certain combinations of options appear together more often than chance would predict. We call this co-selection: the tendency of selections across different systems to attract one another within the same clause, above and beyond what their individual frequencies alone would generate. The degree to which certain systems are mutually attracted – and to which that

attraction is register-specific rather than random – is what we term associative keyness (Figueredo et al., 2024): keyness that is defined not by the frequency of a single system but by its tendency to co-select with others.

To discover recurring co-selection patterns, we applied the FP-growth algorithm (Han, Pei & Yin, 2000) to each sub-corpus independently. FP-growth efficiently enumerates all sets of options – the itemsets – that appear together in at least a minimum number of clauses: set at 5 clauses for Sub-corpora 1 and 2, and 7 for Sub-corpus 3 (which has a wider option space; the higher floor prevents memory overflow without loss of substantive patterns). We analysed in detail only itemsets of 10 or more simultaneous options out of the 19 systems analysed, focusing on multi-system configurations rather than pairwise coincidences.

Three metrics characterise each itemset. Support measures the proportion of clauses in which the full itemset appears (Equation 1). Lift_n asks whether the combination is more frequent than independent co-occurrence would predict: it is the ratio of observed support to the product of each option's individual frequency; lift_n = 1 signals independence, values above 1 signal attraction (Equation 3). Confidence, conf(rest, f), asks: given that all the other options in the itemset are present, how often is option f also present? (Equation 5). Options where conf(rest, f) = 1.000 are invariant members of the core; those with lower confidence are 'loose' members that tend to co-occur with the rest but are not obligatory.

Why association rules are needed, rather than frequency data on their own, is clearest through an example. In Sub-corpus 2 (Informing), both [jussive imperative] MOOD and [no modality] are highly frequent. Because [imperative] clauses cannot select MODALITY – there is no option *[imperative: modality] – both choices are in a sense inevitable together. The question is: which is the defining marker of the Informing register? Frequency alone cannot answer this: both are common. What distinguishes them is their association profile. [jussive imperative] co-selects strongly with obligation modality, addressee subject, and absence of clause-complexing – lift_n = 106.9, meaning the full combination is 106 times more frequent than chance. [No modality], by contrast, carries no such cluster: it simply co-occurs at base rate with whatever other options happen to be present. It is [jussive imperative] that is the key feature; [no modality] is its structural shadow. This distinction is only recoverable from the association metrics, not from frequency data alone.

A cross-analysis then tested whether associative keyness held across the corpus by means of association rules (Hipp et al., 2000). As set out in Section 3.1, the analysis is descriptive: the association rules characterise behaviour observed in the sample, not inferred population probabilities, which is why no inferential tests are applied. Within a fully analysed

sample, ‘frequency’ and ‘probability’ coincide: the probability of an option is its observed rate of selection.

Support measures the proportion of clauses in a corpus C that contain the full itemset (Equation 1):

$$support(IS) = \frac{number\ of\ clauses\ containing\ every\ feature\ in\ IS}{total\ number\ of\ clauses\ in\ the\ sub{-}corpus} \quad (1)$$

An itemset appearing in 38 of Sub-corpus 2’s 252 clauses has a support of 0.151 (Equation 2):

$$support(SC2\ Core\ A) = \frac{38}{252} \approx 0.151\ (15.1\%) \quad (2)$$

lift_n asks whether the itemset is more frequent than independent co-occurrence would predict. The expected support is the product of each option’s individual frequency; lift_n is the ratio of observed to expected (Equation 3). lift_n = 1 signals independence; values above 1 signal attraction:

$$lift\_n(IS) = \frac{support(IS)}{product\ of\ the\ individual\ frequencies\ of\ each\ feature\ in\ IS} \quad (3)$$

For Sub-corpus 2’s core imperative itemset, lift_n = 106.9: the combination is 106 times more frequent than chance (Equation 4):

$$lift\_n(SC2\ Core\ A) = \frac{0.151}{0.405 \times 0.429 \times 0.460 \times 0.448 \times (and\ so\ on)} \approx 106.9 \quad (4)$$

For each option within an itemset, conf(rest, f) asks: given that all the other options are present, how often is f also present? It is the ratio of the full itemset’s support to the support of the set without f (Equation 5):

$$conf(rest, f) = \frac{support\ of\ the\ full\ itemset}{support\ of\ the\ itemset\ without\ f} \quad (5)$$

In Sub-corpus 1 Core A, [circumstantial_1_theme] has conf(rest, f) = 1.000: whenever the other 16 options are present, a circumstantial ideational theme is invariably present (Equation 6). Options with conf < 1.000 are ‘loose’ members – they tend to co-occur with the rest but are not invariant:

$$conf(rest, circumstantial\ theme) = \frac{5}{5} = 1.000 \quad (6)$$

Statistical significance was assessed through a permutation test: the data were shuffled 1,000 times – randomly redistributing which clauses received which value within each system, while preserving total frequencies – and the largest itemset size was recomputed after each shuffle. The p-value is the proportion of shuffles in which the largest shuffled itemset equalled or exceeded the real one. All cores reported in Section 4.3 have p of 0.05 or below; the majority have $p < 0.001$.

3.4 Dynamic Modelling

The dynamic model keeps instantiation and stratification as discrete domains. What is set aside is the stratificational explanation – discourse analysis, functionality, the contextual motivation of the choices: the paper, being driven by instantiation, observes behaviour. Having this in mind, the term ‘genre’ is used in two senses in this paper, and it is important to keep them distinct. In the stratificational sense, a genre is a recurrent configuration of context – a social process with a recognisable staging and purpose (Martin & Rose, 2008). In the instantiational sense, a genre is a sub-population of texts whose systemic behaviour clusters together: not a contextual resource, but a recurrent probability structure. The choice of ‘genre’ for the second sense follows Martin’s (2008, p. 33) use of the term for the semantic potential of instantiation (Figure 4). In what follows, ‘genre’ in the instantiational sense refers to the model derived from the sub-corpus – the probability structure that characterises the sample.

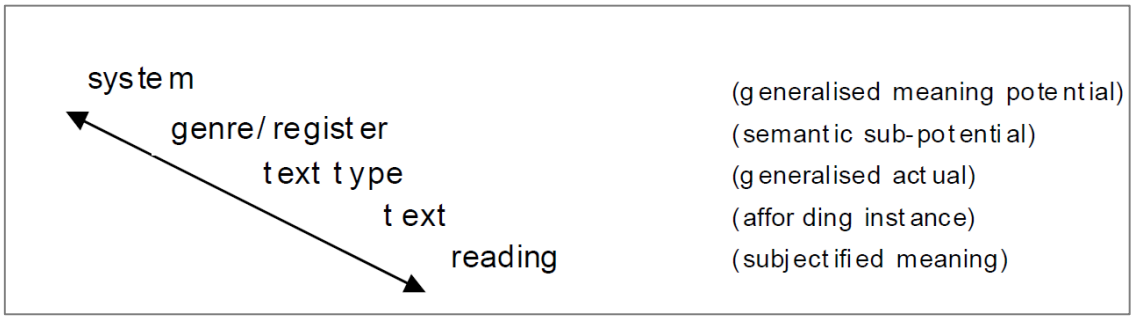

Figure 4 – Genre as a category of instantiation in Martin (2008, p. 33).

Note: Reprinted from Martin (2008, p. 33)

The model treats the clause as a unit of time. A text does not select all its clauses at once; it builds them in sequence, and each clause is a moment at which all 19 systems select together. At some of these moments the register leaves virtually no choice – nearly every text makes the same selection – while at others several options remain open. The dynamic model measures, at each step, how much choice the register leaves available across all systems at once.

To make this measurable, a synthetic text of length L was built for each sub-corpus, where L is the number of clauses in the shortest text in that sub-corpus, so that all five texts contribute data at every clause position. The formal model is set out in the equations that follow; it is then applied to the three sub-corpora in Section 4.4.

Let T be the set of five texts in a sub-corpus, S the set of 19 systems, and L the synthetic-text length. For each clause position i (from 1 to L) and each system s, let O_{is} be the set of features observed across the five texts at that position – where a feature may be a single option or, for recursive systems such as CIRCUMSTANCE, a combination of simultaneous options. The relative frequency of a feature o at position i in system s is the proportion of the five texts that select it (Equation 7):

$$p(o, i, s) = \frac{number\ of\ texts\ selecting\ o\ at\ position\ i\ in\ system\ s}{number\ of\ texts\ in\ T} \quad (7)$$

A position-system cell is fixed when its most frequent feature is selected by at least two thirds of the texts – that is, by at least four of the five. Writing p*(i, s) for the frequency of that most frequent feature, the cell is fixed when p*(i, s) is at least two thirds, and free otherwise (Equation 8):

$$fixed(i, s) = 1\ when\ p*(i, s)\ is\ at\ least\ 2/3;\quad fixed(i, s) = 0\ otherwise \quad (8)$$

The branching factor of a clause position, b at position i, measures how much choice that clause leaves open across all 19 systems at once. A fixed system contributes 1 (no choice); a free system contributes the number of distinct features observed at that position. The branching factor is the product of these contributions over the 19 systems (Equation 9):

$$b(i) = \\ prdct.over\ all\ 19\ systems\text{: }1\ if\ fixed, or\ the\ no.of\ observed\ features\ if\ free \quad (9)$$

The total number of distinct synthetic texts the model admits, N, is the product of the branching factors across all L clause positions. Because this number is very large, it is reported as its base-10 logarithm, which is the sum of the per-position logarithms (Equation 10):

$$N\ =\ product\ b(i)\ over\ positions\ 1\ to\ L;\ log\ (N)\ sum\ log(b(i))\ over\ 1\ to\ L \quad (10)$$

Two points follow from this formulation. First, the branching factor is a property of the clause position, not of any single system: it captures the joint openness of all 19 systems at one moment in the text. A clause where MOOD, MODALITY, and SUBJECT person are all fixed but PROCESS TYPE and the logico-semantic relation are free leaves less open – a smaller branching factor – than a clause where several systems remain undecided. Second, N counts only the configurations the register actually admitted in the five texts, not every logically possible combination. A synthetic text is admissible when each of its clauses is attested somewhere in the sub-corpus; N is the number of distinct texts of length L that can be built this way.

Because N grows with the length L, the raw totals are not directly comparable across sub-corpora of different lengths. The comparable figure is the per-position branching – log N divided by L – which gives the average amount of choice the register leaves open at each clause. The results of applying this model to the three sub-corpora are reported in Section 4.4.

4. Results

4.1 Systemic Profiles of the Sub-corpora

Read instantiationally, the overall probability configuration corresponding to a genre is reflected in which options dominate at the level of the sub-corpus as a whole. Table 3 reports, for each system, its dominant option in each sub-corpus and that option's relative frequency. The denominator for each system is the number of clauses in which it applies (non-selectable and blank cells excluded), stated where it departs from the sub-corpus total; cells where the dominant option itself shifts across sub-corpora are marked with an asterisk. The table shows which systems behave similarly across registers (the para-redundant backbone) and which shift

markedly (the redistribution carriers). For example, a quick scan shows that POLARITY, THEME: interpersonal, and ENGAGEMENT are stable across all three sub-corpora – high proportions of [positive], [no interpersonal theme], and [monoglossia] in all three. These are the systems that hold steady regardless of genre family. By contrast, MOOD, MODALITY, SUBJECT: person, and DEIXIS shift substantially, and it is in these systems that the three types diverge.

Table 3 – Systemic profile of the sub-corpora (dominant option and relative frequency)

System	Sub 1 (Evaluating)	Sub 2 (Informing)	Sub 3 (Engaging)
MOOD	ind. declarative 94.3%	ind. declarative 55.2%	ind. declarative 91.7%
POLARITY	positive 98.1%	positive 93.7%	positive 89.6%
DEIXIS	present 60.4%	present 85.3%	*past 52.1% (n=288)
PROCESS TYPE	material: transf. 34.6%	material: transf. 55.6%	material: transf. 40.8%
MODALITY	no modality 87.4%	no modality 47.6%	no modality 90.7%
THEME: textual	no textual 58.5%	no textual 75.0%	no textual 62.3%
THEME: interpersonal	no interp. 93.7%	no interp. 94.0%	no interp. 92.0%
THEME: ideational	no ideational 77.4%	no ideational 94.0%	no ideational 85.5%
CIRCUMSTANCE	no circ. 52.8%	no circ. 61.9%	no circ. 63.3%
CONNECTION	no connection 71.1%	no connection 52.0%	no connection 61.2%
LOGICO-SEMANTIC	*elaboration 30.2%	*no selection 42.9%	*extension 43.6%
TAXIS: parataxis	no parataxis 37.7%	no parataxis 61.9%	*parataxe 1 28.4%
TAXIS: hypotaxis	no hypotaxis 78.0%	no hypotaxis 66.3%	no hypotaxis 87.9%
SUBJECT: person	non-interactant 100% (n=150)	non-interactant 52.7% (n=239)	*speaker 1P 55.0% (n=282)
SUBJECT: responsibility	responsible 89.9% (n=148)	responsible 96.7% (n=244)	responsible 98.6% (n=282)
SUBJECT: number	singular 69.0% (n=145)	singular 76.2% (n=240)	singular 91.8% (n=282)
SUBJECT: presumption	explicit 88.4% (n=147)	*implicit 53.6% (n=239)	explicit 53.2% (n=284)
MESSAGE: periodicity	HyperTheme 64.2%	HyperTheme 3 25.4%	HyperTheme 1 57.8%
ENGAGEMENT	monoglossia 77.4%	monoglossia 84.9%	monoglossia 79.2%

Five further systems were held constant within each sub-corpus and so are not shown; an invariant system cannot enter a pattern of variation and was excluded from the co-selection analysis. In Sub-corpus 1 these are MODE (monologic reflection), TENOR status (equal), TENOR solidarity (distant), SPEECH FUNCTION (giving information), and EXCHANGE (initiating move); the same five are constant in Sub-corpora 2 and 3, with values appropriate to each.

Examples (2)–(4) present one typical clause from each sub-corpus. Each annotation names the dominant options from Table 3 and shows how they converge in a single clause, instantiating the type profile.

(2) Sub-corpus 1 – Evaluating [EVA_05, clause 1]:

No ano de 2013,	*aproximadamente 382 milhões de pessoas no mundo*	*possuíam*	*essa doença.*
in.the year of 2013	approximately 382 million of people in.the world	possessed.3PL	this disease

'In 2013, approximately 382 million people in the world had this disease.'

MOOD: [indicative declarative] (94.3% in Sub 1); SUBJECT person: [non-interactant] (100%); MODALITY: [none] (87.4%); DEIXIS: [past] (28.3%); ENGAGEMENT: [monoglossia] (77.4%). From an stratificational point of view, we can say that this clause configuration maps onto a biomedical fact about a population at a remove from both the writer and the reader – the defining orientation of the Evaluating genre.

(3) Sub-corpus 2 – Informing [DIR_04, clause 30]:

Diversifique	*a*	*alimentação*	*para*	*um*	*bom*	*funcionamento*	*do*	*organismo!*
diversify.IMP.2SG	the	diet	for	a	good	functioning	of.the	organism

'Diversify your diet for the good functioning of the organism!'

MOOD: [jussive imperative] (40.5% in Sub 2, 0% in Sub 1 and ~2% in Sub 3); SUBJECT person: [addressee, 2nd person] (46.0%); MODALITY: [obligation] (44.8%); PROCESS TYPE: [material transformative] (55.6%). This configuration maps onto what stratificationally can be described as concentrating the interpersonal resources of Informing: the reader is cast as the agent of a required action.

(4) Sub-corpus 3 – Engaging [ENG_05, clause 23]:

Demos	*largada*	*às*	*04:30*	*da*	*manhã.*
gave.1PL	start	at.the	04:30	of.the	morning

'We set off at 4:30 in the morning.'

MOOD: [indicative declarative] (91.7%); SUBJECT person: [speaker, 1st person] (55.0%); DEIXIS: [past] (52.1%); LOGICO-SEMANTIC: [extension] (43.6%); MODALITY: [none] (90.7%). Stratificationally, it can be said that the speaker is the protagonist of their own narrative. The temporal anchor is the past; the relation to the preceding clause is extension – a further event added to the chain.

The three clauses trace a clean instantiational path: a third-person declaration of biomedical fact (Evaluating), a second-person directive obligation (Informing), and a first-person narrative event (Engaging). Same resource; three non-overlapping configurations.

4.2 Redistribution

Redistribution is the shift in an option's relative frequency from one sub-corpus to another, with Sub-corpus 1 (Evaluating) as the prior. Table 4 isolates the most telling options – those whose frequency moves most substantially across the three registers. Reading the table column by column reveals the profile of each type; reading it row by row reveals the trajectory of each option across the cline. Options that are absent from one column and dominant in another are the clearest redistribution carriers; options that hold similar values across columns belong to the para-redundant backbone.

A received way of reading such profiles is to set them against the behaviour of language in general. Halliday and James (1993) proposed that linguistic systems fall into two types: equiprobable and skewed. On that approach, the present data can be compared against language in general as well – a comparison we return to, and qualify, in Section 4.5.

Table 4 – Selected redistributions across the sub-corpora (Sub 1 = prior)

System	**Feature**	**Sub 1 (Evaluating)**	**Sub 2 (Informing)**	**Sub 3 (Engaging)**
MOOD	indicative declarative	94.3%	55.2%	91.7%
MOOD	jussive imperative	0% (absent)	40.5%	~2%
MODALITY	no modality	87.4%	47.6%	90.7%
MODALITY	obligation	~0%	44.8%	~0%
PROCESS TYPE	material: transformative	34.6%	55.6%	40.8%
SUBJECT: person	non-interactant (3rd)	100%	52.7%	44.0%
SUBJECT: person	addressee (2nd)	0% (absent)	46.0%	~1%
SUBJECT: person	speaker (1st)	0% (absent)	0% (absent)	55.0%
DEIXIS	present	60.4%	85.3%	45.1%
DEIXIS	past	28.3%	~2%	52.1%
LOGICO-SEMANTIC	expansion: extension	n.r.	n.r.	43.6%

LOGICO-SEMANTIC	expansion: elaboration	30.2%	24.2%	26.0%
POLARITY	negative	1.9%	6.3%	10.4%
SUBJECT: number	singular	69.0%	76.2%	91.8%

Note: n.r. = not separately reported (the option falls below the two dominant options in that sub-corpus's profile). DEIXIS and SUBJECT percentages are computed over the clauses where each system applies (e.g. Sub 2 DEIXIS excludes the 102 imperative clauses), so denominators differ across sub-corpora; see Table 3.

The most consequential redistribution in this corpus is not a simple monotonic cline. [jussive imperative] and [modality: obligation] both peak sharply in the Informing genre family (40.5% and 44.8% respectively) and collapse to near-zero in both the Evaluating and Engaging families – they bracket, rather than cline. Sub-corpus 2 is the most marked: it carries configurations absent from both flanking genre families.

Examples (2)–(4) above illustrate this directly. Example (2) (Evaluating) is a positive declarative with no modality – MOOD: declarative (94.3%), MODALITY: none (87.4%), SUBJECT: non-interactant. Example (3) (Informing) is a jussive imperative with obligation modality – MOOD: jussive imperative (40.5%), MODALITY: obligation (44.8%), SUBJECT: addressee 2nd person (46.0%). Example (4) (Engaging) returns to declarative with no modality – but with SUBJECT: speaker 1st person (55.0%) and DEIXIS: past (52.1%). The same grammatical bracket (declarative, no modality) at the Evaluating and Engaging ends; an entirely different social orientation: the Evaluating clause asserts a fact about a population; the Engaging clause narrates a personal event. The Informing clause between them is the grammatical outlier.

The cleanest monotonic redistribution is in SUBJECT person: non-interactant declines from 100% to 52.7% to 44.0%, while speaker (first person) rises from zero to zero to 55.0%, and addressee (second person) rises from zero to 46.0% then falls back to near-zero. Three genre families; three distinct person configurations. Deixis follows similarly: present tense dominates in the Evaluating and Informing registers (60% and 85%), while past tense rises to 52.1% in the Engaging narrative – the temporal anchor of the recount.

Negative polarity increases monotonically (1.9%, then 6.3%, then 10.4%), and singular number likewise (69%, then 76%, then 92%), the latter reflecting the increasingly individual orientation of the Engaging register. Material process redistributes in an arc (34.6%, then 55.6%, then 40.8%), peaking at Informing where the instructional character foregrounds doing. The discourse systems – theme, circumstance, connection – remain relatively stable across all three registers, suggesting they belong to the para-redundant backbone rather than the redistribution carriers.

4.3 Associative Keyness

The second mechanism of instantiation is co-selection: the tendency of certain options across different systems to attract one another within the same clause, beyond what their individual frequencies predict. Co-selection was operationalised through the FP-growth itemset analysis described in Section 3.3. For each sub-corpus, the largest itemset in the real data was compared against 1,000 permutations; all reported cores have p of 0.05 or below, and the great majority have $p < 0.001$.

The analysis returns co-selection 'cores': itemsets of 10 or more simultaneous options that recur in multiple clauses well above chance. Sub-corpus 1 has two cores, Sub-corpus 2 has three, and Sub-corpus 3 has three. Table 5 summarises the cores with their key metrics and most identifying options – those with the highest lift contribution, i.e. the options one would least expect to find in the combination given the others. The table should be read alongside the metrics defined in Section 3.3: n is the size of the itemset (number of simultaneous options), Support is the proportion of sub-corpus clauses in which the full core appears, lift_n is how many times more frequent the combination is than chance predicts, and the identifying options are those with the highest conf(rest, f). A high lift_n with a small number of clauses (as in SC1 Core A) indicates a tightly bound, rare configuration; a lower lift_n with many clauses (as in SC2 Core A) indicates a pervasive, moderate-strength cluster.

Table 5 – Co-selection cores by sub-corpus

Sub-corpus	Core	n	Clauses	Support	lift_n	p	Key identifying options (lift contribution)
SC1 Evaluating	A	17	5	0.031	141.4	<0.001	circ_1_theme (4.97); no_logico_semantic (4.68)
SC1 Evaluating	B	15	9	0.057	48.2	<0.001	relative_theme (5.68); parataxis_2 (3.49); elaboration (2.71)
SC2 Informing	A	15	38	0.151	106.9	<0.001	jussive_imperative (2.47); no_logico_semantic (2.33); addressee_2P (2.29); obligation_modality (2.23)
SC2 Informing	B	~15	–	–	150.8	<0.001	SC2 Core A nucleus + additive connection
SC2 Informing	C	~15	–	–	133.4	<0.001	SC2 Core A nucleus + additive + material process
SC3 Engaging	A	15	12	0.042	4.9	<0.001	recoverable_implicit (1.89); speaker_1P (1.86); extension (1.53)
SC3 Engaging	C	15	11	0.038	12.1	<0.001	additive (2.75); extension (2.29); recoverable_implicit (2.02); speaker_1P (1.86)

The patterns emerging are as follows. Sub-corpus 2 (Informing) shows the strongest co-selection by frequency: Core A alone covers 38 clauses (15.1% of the sub-corpus) at lift_n = 106.9. The invariant nucleus of all three SC2 cores – jussive imperative + obligation modality + second-person addressee subject + no logico-semantic relation – is 106 times more frequent than independent co-occurrence would predict. Despite both being highly frequent options in SC2, jussive imperative and material process do not form a discovery: their lift_n ≈ 1 (they co-occur at base rate). High frequency does not equal co-selection. The Informing text is defined not by the combination of imperative and material, but by the full nucleus: directive mood + obligation + reader-as-agent + no clause-complexing.

Second, Sub-corpus 1 (Evaluating) has the tightest individual binding: Core A (lift_n = 141.4) is the most statistically unlikely combination, though it covers only 5 clauses. Whenever a clause in the academic article has a circumstantial ideational theme – a marked, fronted circumstantial – it is invariably also not in a clause-complex (conf = 1.000 for both identifying options). This is the Evaluating register's rare but tightly bound configuration: the fronted-circumstance, stand-alone assertion.

Third, Sub-corpus 3's (Engaging) invariant nucleus – speaker_informal_1P + recoverable_implicit + extension – is the major pattern of the patient statement: first-person narrator, implicit subject (I/we), chaining by extension (and/but). Core C adds additive connection, confirming that the narrative's clause-nexusing is the defining co-selection pattern – not what the clauses say about diabetes, but how they are strung together in the first person.

Example (4) above – 'Demos largada às 04:30 da manhã' ('We set off at 4:30 in the morning') – is an attested instance of SC3 Core C. Its annotation confirms the invariant nucleus: SUBJECT:person = speaker_informal_1P, SUBJECT:presumption = recoverable_implicit, LOGICO-SEMANTIC = extension. The additive connection (CONNECTION = additive) that Core C adds over Core A is also present: the clause chains onto the previous event with 'and' (implicit in Portuguese), adding one more step to the personal recount.

The second mechanism of instantiation is the keyness given by co-selection: multi-system configurations recur far above chance, in ways that are register-specific (SC2's directive nucleus, SC3's narrative chaining) yet partly shared (the imperative–material independence, confirming that frequency and association are orthogonal measures).

4.4 Dynamics

The profile and co-selection analyses treat the clause as a unit of simultaneous selection, but set aside the order in which clauses accumulate into a text. The third mechanism of instantiation – dynamics – reinstates time: it asks how much choice the text type leaves open as the text unfolds, clause by clause. The dynamic model defined in Section 3.4 was applied to each sub-corpus.

For each sub-corpus, every clause position was classified system by system as fixed or free under the two-thirds norm. Tables 6–8 report, for each sub-corpus and each of the 19 systems, the number of clause positions (out of L) at which that system is fixed. The bottom row of each table gives the total proportion of fixed cells and the number of distinct synthetic texts the model admits, reported as a power of 10.

How to read these tables: a system with a high fixed count is one the text type settles early and holds constant – it offers the text-maker little choice. A system with a low fixed count stays open across the text, carrying most of the variation between one text and another. The admissible-text figure at the foot of each table is the total number of distinct texts of length L that the register permits; the larger it is, the more room the text type leaves for individual texts to differ.

Table 6 – Dynamic model: fixed positions per system (Sub 1, Evaluating, L = 24 clauses)

System	**Fixed**	**System**	**Fixed**
MOOD	23	TAXIS: hypotaxis	18
POLARITY	24	SUBJECT: person	22
DEIXIS	5	SUBJECT: responsibility	17
PROCESS TYPE	3	SUBJECT: number	12
MODALITY	20	SUBJECT: presumption	19
THEME: textual	8	MESSAGE: periodicity	20
THEME: interpersonal	21	ENGAGEMENT	17
THEME: ideational	19		
CIRCUMSTANCE	5		
CONNECTION	13		
LOGICO-SEMANTIC	2		
TAXIS: parataxis	1		
Total fixed	**269 / 456**	**Admissible texts**	**approx. 10^77.9**

Table 7 – Dynamic model: fixed positions per system (Sub 2, Informing, L = 41 clauses)

System	**Fixed**	**System**	**Fixed**
MOOD	20/41	TAXIS: hypotaxis	21/41

POLARITY	39/41	SUBJECT: person	16/41
DEIXIS	11/39	SUBJECT: responsibility	40/41
PROCESS TYPE	13/41	SUBJECT: number	24/41
MODALITY	19/41	SUBJECT: presumption	21/41
THEME: textual	25/41	MESSAGE: periodicity	14/41
THEME: interpersonal	41/41	ENGAGEMENT	34/41
THEME: ideational	40/41		
CIRCUMSTANCE	15/41		
CONNECTION	14/41		
LOGICO-SEMANTIC	4/41		
TAXIS: parataxis	14/41		
Total fixed	**425 / 777**	**Admissible texts**	**approx. 10^144.4**

Table 8 – Dynamic model: fixed positions per system (Sub 3, Engaging, L = 39 clauses)

System	**Fixed**	**System**	**Fixed**
MOOD	39/39	TAXIS: hypotaxis	35/39
POLARITY	34/39	SUBJECT: person	11/39
DEIXIS	14/39	SUBJECT: responsibility	38/39
PROCESS TYPE	2/39	SUBJECT: number	35/39
MODALITY	34/39	SUBJECT: presumption	11/39
THEME: textual	20/39	MESSAGE: periodicity	14/39
THEME: interpersonal	37/39	ENGAGEMENT	29/39
THEME: ideational	32/39		
CIRCUMSTANCE	15/39		
CONNECTION	15/39		
LOGICO-SEMANTIC	8/39		
TAXIS: parataxis	4/39		
Total fixed	**427 / 741**	**Admissible texts**	**approx. 10^134.0**

Note on cross-corpus comparison: the raw exponents (10^77.9, 10^144.4, 10^134.0) scale with synthetic-text length (L = 24, 41, 39 respectively) and are not directly comparable. Per-clause-position branching is more informative: ~10^3.2 per position for Sub 1, ~10^3.5 for Sub 2, ~10^3.4 for Sub 3 – Sub-corpus 2 (Informing) admits the most variation per clause position despite having the highest proportion of fixed cells.

Table 9 tells a complementary story. Each row of Table 9 is a distinct co-selection configuration – a clause type. If every clause selected differently, the number of distinct types would equal the number of clauses. In practice, some types recur. The question is how many. A register with strong co-selection pressure (many fixed positions) would be expected to repeat the same clause types frequently; one with low pressure would show near-maximum dispersion. Table 9 shows that all three sub-corpora are highly dispersed, but for different reasons.

Table 9 – Clause-type dispersion across the sub-corpora

	Sub 1 – Articles	**Sub 2 – Booklets**	**Sub 3 – Statements**
Total clauses	159	252	289
Distinct clause types	154 (97.0%)	229 (90.9%)	284 (98.3%)
Most frequent type	2 occ. (1.3%)	8 occ. (3.2%)	3 occ. (1.0%)
Types to cover 67%	102	146	189

The Engaging patient narrative (Sub 3) is the most dispersed of the three registers: 284 distinct clause types in 289 clauses, with no type covering more than 1% of the corpus. This is at first surprising, since Sub-corpus 3's dynamic model fixes a high proportion of positions (427/741 = 57.6%). The apparent contradiction resolves when the fixed positions are examined by system. In Sub 3, the it is concentrated in the grammatical backbone: MOOD is fully determined (39/39 positions fixed at declarative), modality fixed (34/39), hypotaxis fixed (35/39), responsibility fixed (38/39), and number fixed (35/39). But the systems that carry content variation – PROCESS TYPE (2/39 fixed), TAXIS parataxis (4/39), SUBJECT person (11/39), SUBJECT presumption (11/39), and LOGICO-SEMANTIC (8/39) – remain highly variable. The personal narrative is grammatically uniform but experientially diverse: every story follows the same interpersonal and grammatical template, but the world it reports differs clause by clause. Sub-corpus 2 is the inverse: more associations at the genre level but more per-position branching, because the instructional structure itself generates clause-complexing variation. Sub-corpus 1 is the most under-determined at the grammatical backbone level (PROCESS TYPE 3/24, LOGICO-SEMANTIC 2/24, TAXIS parataxis 1/24) because the expository article does not privilege any particular experiential configuration.

This is the third mechanism of instantiation: dynamics determines what kind of variation is possible as the text develops, and that variation is inversely proportional to keyness – but keyness itself is system-specific, so the genre's grip may be tight at the interpersonal and textual level while remaining loose at the experiential level, or vice versa.

4.5 Mechanisms of Instantiation and Their Implications for Stratification

Within the architecture of the system, the three mechanisms specific to instantiation – redistribution, association, and dynamics – relate to all strata. From above, register fixes a range of possible configurations for the probabilistic realisation of a genre. The way a difference in genre family plays out across the grammar is visible in the redistribution of MOOD across the three sub-corpora.

Consider the realisation of a directive or evaluative move across the three genre families, illustrated by examples (2)–(3) above and by the profile data in Table 4:

(5) [EVA_05, cl.1, Evaluating – declarative assertion, 3rd person, past]:
'In 2013, approximately 382 million people in the world had this disease.'
MOOD: [indicative declarative] (94.3%); SUBJECT: [non-interactant] (100%); MODALITY: [none] (87.4%).

(6) [DIR_04, cl.30, Informing – jussive imperative directive, 2nd person, obligation]:
'Diversify your diet for the good functioning of the organism!'
MOOD: [jussive imperative] (40.5%); SUBJECT: [addressee, 2nd person] (46.0%); MODALITY: [obligation] (44.8%).

(7) [ENG_05, cl.23, Engaging – declarative narration, 1st person, past]:
'We set off at 4:30 in the morning.'
MOOD: [indicative declarative] (91.7%); SUBJECT: [speaker, 1st person] (55.0%); MODALITY: [none] (90.7%).

The MOOD data shows that the redistribution is not a gradient: jussive imperative does not increase steadily from Evaluating to Engaging. Instead, it brackets – rising from zero in Evaluating to 40.5% in Informing, then collapsing back to near-zero in Engaging. Obligation modality follows the same pattern. The Informing genre family is where the system concentrates directive force; neither the Evaluating nor the Engaging family calls on the imperative in this domain. The Engaging patient narrative returns to declarative because its move is not instruction but testimony – a different discourse function, not a weakened version of the same one.

The SUBJECT person redistribution is the clearest three-way differentiator. The Evaluating article speaks of diabetes in the third person, positioning knowledge objects rather than interactants. The Informing booklet splits nearly evenly between third person and second person (addressee) – the register alternates between presenting information and directly instructing the reader. The Engaging statement is anchored in the first person: the speaker is their own topic. This three-way configuration – third person, then a second/third split, then first person – is not derivable from any single system; it emerges from the joint redistribution of SUBJECT person, MOOD, MODALITY, and DEIXIS taken together.

The three clauses in hand trace this configuration exactly. Example (3): SUBJECT:person = [non_interactant], MOOD = [declarative], DEIXIS = [past], MODALITY = [none] – the disease as object of knowledge. Example (4): SUBJECT:person = [addressee_informal_2P], MOOD = [jussive_imperative], MODALITY = [obligation] – the reader as agent of required action. Example (4): SUBJECT:person = [speaker_informal_1P], MOOD = [declarative], DEIXIS = [past], MODALITY = [none] – the speaker as protagonist of their own narrative. Same three systems; three configurations that do not overlap.

Associative keyness shows the same cross-stratal logic. One coupling – intensification with hypotaxis (the clause-complexing backbone) – recurs in all three sub-corpora: the shared spine of clause-nexusing, independent of genre. The other strong associations are register-specific: the conditional-logic cluster (conditional connection, intensification, and hypotaxis together) is confined to Sub-corpus 2; the affective-stance cluster (affect together with desiderative mental process) is confined to Sub-corpus 3; and the coupling of heteroglossic projection with verbal process appears in Sub-corpus 1 and recurs in Sub-corpus 3, because both the expository article and the patient narrative cite and report – though from opposite vantage points. From the perspective of instantiation, these associations are not properties of the systems but of how instances accumulate across texts: they emerge from counting.

The most important contribution of the mechanisms of instantiation to the relationship with stratification is this: all strata instantiate (Martin, 2009). From stratification, this means the strata cannot by themselves yield a specific text; what they do is abstract an organisation of ‘patterns of patterns’ so that any instance contains all patterns from expression to context. From instantiation, what exist are more probable associations among systems, distributed in distinct ways; these associations in turn generate the strata, which can be understood as types of association among systems. The phenomena Martin (2008) identifies as ‘coupling’ and ‘commitment’ are exemplary cases of this process.

A final implication concerns how redistribution is to be measured. The three sub-corpora differ in genre family, so the profile shifts in Tables 2 and 3 are attributable to genre-family membership, not to differences in subject matter. This is why redistribution is read sample-to-sample – each sub-corpus against the Sub 1 prior – rather than against a global profile for the language. Instantiation redistributes the behaviour of a population over the same resource without touching the resource itself; a global, resource-level probability is therefore not the baseline for redistribution. This places the equiprobable comparison of Section 4.2 in a particular light: the equiprobable test asks whether a system is skewed at all within a sample; redistribution asks how a profile shifts from one sample to another. The present data bear directly on the second.

Conclusion

The paper set out to show how a quantitative description feeds the view of text as a unit of probability, and to define text as a unit of variation in the take-up of linguistic resources. To that end, it examined the linguistic configuration of three sub-corpora – matched in field and stepped across three genre families – to observe the principles of instantiation and to identify possible mechanisms of its operation.

Three mechanisms of instantiation emerged through association analysis and dynamic modelling. Redistribution takes up resources in biased proportions across the sub-corpora, since no system is selected equiprobably in any actual instance of use. Association probabilistically relates choices across systems, producing patterns of correlation; because these patterns are themselves redistributed, they give rise to keyness, and key systems are fundamental to the separation of the potential into registers and of registers into types of text. Dynamics determines how redistribution and association are arranged across the temporal development of the text.

The results show that probability is a property of instantiation, not a product of any ‘intentional random function’ of the linguistic system. It follows from the sheer number of configurations a text admits: the Sub-1 Evaluating model alone admits approximately 10^77.9 distinct synthetic texts at 24 clause positions. The chance of two texts in this domain coming out exactly alike is effectively nil – a figure that becomes meaningful only in contrast: Sub-corpus 3’s 289 clauses produce 284 distinct clause types, while Sub-corpus 2’s stronger Informing constraint still yields 229 distinct types in 252 clauses.

The results also indicate some of the mechanisms that contribute to the semantic unification of a text from the standpoint of instantiation. Where stratification defines 'text' abstractly as a kind of 'macrostructure' that realises co-selection, instantiation defines 'text' as a configuration of linguistic units, probabilistically distributed, with key systems fixed by association, that unfolds in time.

On this basis, instantiation may be reinterpreted as the expression of the dynamic probability of the text (Peixoto et al., 2026). Viewed from instantiation, what exists is a profusion of instances that vary without obeying any pre-established probabilistic norm. Instantiation distributes and redistributes resources, builds associations, and unfolds them in time; it does not, however, determine which redistributions and associations come to count as models – that is the work of stratification, one that we can only find when theory and description are in place.

Two limitations bound these findings. The results hold for the research sample; the mechanisms identified may operate differently in other text types, or may not surface at all. And the question of why it is stratification, or individuation for that matter, that fixes which models consolidate and come to be replicated – whether through chance, the relation between language and society, or ideology – lies beyond what instantiation can answer.

These limits also mark the way forward. A fuller account would recast the instantiation cline itself as a relation between sample and population, and would ask whether the 'global probabilities' attributed to the language are a property of the resource or the reified behaviour of a single sample. That reframing, and its consequences for sub-potentialisation, ought to be taken up in subsequent work.

Conflito de interesses

Os autores declaram que não existe conflito de interesses de qualquer ordem que possa influenciar esta pesquisa.

Contribuição de autoria

Giacomo Figueredo: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, writing.

Francieli Oliveira: conceptualization, data curation, formal analysis, investigation, methodology, writing.

Eduardo Luz: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, software, validation, writing.

Disponibilidade de dados

Os conjuntos de dados brutos, corpora analisados e planilhas de análise estão prontamente disponíveis mediante solicitação aos autores.

Agradecimentos

Agradecemos o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq no financiamento a esta pesquisa.

References

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Christie, F., & Martin, J. R. (1997). *Genre and institutions: social processes in the workplace and school*. Cassel.

Doran, Y. J. (in prep.) *Modelling context from multiple perspectives*.

Doran, Y. J., Martin, J. R., & Herrington, M. (2024). *Rethinking context: realisation, instantiation, and individuation in systemic functional linguistics*. *Journal of World Languages*, *10*(1), 177–220. https://doi.org/10.1515/jwl-2023-0051

Ellis, J., & Ure, J. (1969). Language varieties: register. In A. R. Meetham & R. Hudson (Eds.), *Encyclopaedia of Linguistics, Information, and Control* (pp. 251–289). Pergamon Press.

Figueredo, G. (2007). *Uma descrição sistêmico-funcional da estrutura do grupo nominal em português orientada para os estudos linguísticos dos estudos multilíngues* (Dissertação de Mestrado). PosLin/UFMG, Belo Horizonte.

Figueredo, G. (2011). *Introdução ao perfil metafuncional do português brasileiro*: contribuições para os estudos multilíngues (Tese de Doutorado). PosLin/UFMG, Belo Horizonte.

Figueredo, G. (2020). Axial Argumentation and Cryptogrammar in Textual Grammar: THEME in Brazilian Portuguese. In J. R. Martin, Y. J. Doran, & G. Figueredo (Eds.), *Systemic Functional Language Description - Making Meaning Matter* (1st ed., pp. 129–161). Routledge.

Figueredo, G. (2021a). *Interpersonal Grammar in Brazilian Portuguese* (1st ed.). Cambridge University Press.

Figueredo, G. (2021b). *The nominal group in Brazilian Portuguese*. Word – Journal Of The International Linguistic Association, 67, 461–492.

Figueredo, G., Santorum, K. A. T., Andrade, N. B., & Damasceno, L. A. (2022). Modelagem de produção de texto sintético como subsídio para a aplicação embasada na Pedagogia de Gênero / Modelling Text Synthesis as a Basis for Application of Genre Pedagogy. *REVISTA de ESTUDOS DA LINGUAGEM*, *30*(2), 700. https://doi.org/10.17851/2237-2083.30.2.700-742

Figueredo, G., Freitas, G. G. B., Azevedo, L. S., & Damasceno, L. A. (2024). Considerações sobre a organização do texto e da instanciação sob a perspectiva sistêmico-funcional. *DELTA: Documentação de Estudos Em Lingüística Teórica E Aplicada*, *40*(1). https://doi.org/10.1590/1678-460x202440159824

Figueredo, G., Pagano, A., & Ferreguetti, K. (2014). Os sistemas textuais de focalização na organização funcional da gramática do português brasileiro. D.E.L.T.A.: Documentação de Estudos Em Lingüística Teórica E Aplicada, 30(2), 309–352.

Figueredo, G., & Freitas, G. (2026). The verbal group system and structure in Brazilian Portuguese. *Lingua*, 329, 104070. https://doi.org/10.1016/j.lingua.2025.104070

Firth, J. R. (1957). Modes of Meaning. In F. R. Palmer (Ed.), *Papers in Linguistics.* Oxford University Press.

Halliday, M. A. K. (1978). *Language as Social Semiotic: The social interpretation of language and meaning* (pp. 20–38). Edward Arnold & University Park Press.

Halliday, M. A. K. (1991). Towards probabilistic interpretations. In E. Ventola (Ed.), *Functional and Systemic Linguistics: approaches and uses* (pp. 39–61). De Gruyter Mouton. https://doi.org/10.1515/9783110883527

Halliday, M. A. K. (1992). How do you mean?. In M. Davies & L. Ravelli (Eds.), *Advances in Systemic Linguistics: Recent Theory and Practice* (pp. 20–25). Pinter Publishers.

Halliday, M. A. K. (2005). On matter and meaning: the two realms of human experience. *Linguistics and the Human Sciences*, *1*(1). https://doi.org/10.1558/lhs.v1i1.59

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Halliday, M. A. K., & JAMES, Z. L. A. (1993). A quantitative study of polarity and primary tense in the English finite clause. In J. M. Sinclair, M. Hoey, & G. Fox (Eds.), *Techniques of description: spoken and written discourse (A festschrift for Malcolm Coulthard)* (pp. 32–66). Routledge.

Halliday, M. A. K., & Greaves, W. S. (2008). *Intonation in the Grammar of English*. Equinox Publishing.

Hao, J., & Wang, P. (2022). Chinese Nominal Groups: The Metaphorical Realization of Figures. *WORD* 68(2):109-144. https://doi.org/10.1080/00437956.2021.2024352

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining: a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, *2*(1), 58–64. https://doi.org/10.1145/360402.360421

Lemke, J. (1991). Text production and dynamic text semantics. In E. Ventola (Ed.), *Functional and systemic linguistics: approaches and uses*. Mouton de Gruyter.

Lemke, J. (1995). Intertextuality and text semantics. In P. Fries, & M. Gregory (Eds.), *Discourse in society: Systemic functional perspectives*. Ablex Publising.

Martin, J. R. (1992). *English text: system and structure*. John Benjamins. https://doi.org/10.1075/z.59

Martin, J. R. (2008). Tenderness: Realisation and instantiation in a Botswanan Town. *Odense Working Papers in Language and Communication*, *29*, 30–58.

Martin, J. R. (2009). Realisation, instantiation and individuation: some thoughts on identity in youth justice conferencing. *DELTA: Documentação de Estudos Em Lingüística Teórica E Aplicada*, *25*(spe), 549–583. https://doi.org/10.1590/s0102-44502009000300002

Martin, J. R. (2010). Semantic Variation – Modelling Realisation, Instantiation and Individuation in Social Semiosis. In Monika Bednarek & J. R. Martin (eds.), *New discourse on language: Functional perspectives on multimodality, identity, and affiliation.* London: Continuum. 1-34.

Martin, J. R. (2015). Meaning beyond the clause: Co-textual relations. *Linguistics and the Human Sciences*, *1*(2-3), 203–235. https://doi.org/10.1558/lhs.34711

Martin, J. R., & Matthiessen, C. M. I. M. (1992). Systemic typology and topology. In F. Christie (Ed.), *Literacy in social processes: papers from the Inaugural Australian Systemic Functional Linguistics Conference*. Deakin University. Reprinted in Martin, J.R. (2010), SFL theory, Volume 1 in The collected works of J.R. Martin, edited by Wang Zhenhua. Shanghai: Shanghai Jiao Tong University Press. 167–215.

Martin, J. R., & Rose, D. (2008). *Genre relations: Mapping culture*. Equinox Pub.

Maton, K. (2014). *Knowledge & knowers: towards a realist sociology of education*. Routledge.

Matthiessen, C. M. I. 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). Pinter.

Matthiessen, C. M. (2015a). Register in the round: registerial cartography. *Springer*, *1*(48), 203–235. https://doi.org/10.1186/s40554-015-0015-8

Matthiessen, Christian M.I.M. 2015b. Modelling context and register: the long-term project of registerial cartography. In *Estudos sistêmico-funcionais: desdobramentos e interfaces*. Revista Letras, Santa Maria, n. 50, jun. 15-90.

Nesbitt, C., & Plum, G. (1988). Probabilities in a systemic-functional grammar: The clause complex in English. *New Developments in Systemic Linguistics*, *2*, 6–38.

Neumann, S. (2013). *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. De Gruyter Mouton.

Neumann, S. (2015). Applying register analysis to varieties of English. In M. Fludernik & B. Kohlmann (Eds.), *Anglistentag 2011 Freiburg Proceedings* (pp. 75–94). Wissenschaftlicher Buchverlag Trier.

Oliveira, F. S. (2022). *O discurso do autocuidado em saúde: uma descrição de gêneros na covariação experto-leigo* (pp. 1–228) [Dissertation]. https://repositorio.ufmg.br/items/5c525955-865b-4680-bce9-875ed33f6ea4

Oliveira, F. S., & Figueredo, G. P. (2020). Análise contrastiva de gêneros textuais como contribuição para o estudo do produto tradutório. *Cadernos de Tradução*, *40*(2), 221–251. https://doi.org/10.5007/2175-7968.2020v40n2p221

Painter, C. (1999). *Learning through language in early childhood*. Cassel.

Peixoto, M., Silva, G., Figueredo, G., Silva, P., & Luz, E. J. S. (2026). Global vs. Local Sentence Embeddings for Brazilian Portuguese: Revisiting Monolingual Models in the Age of Foundation Models. *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026), 1*, 529–539. https://aclanthology.org/2026.propor-1.52/

Plum, G., & Cowling, A. (1987). Social constraints on grammatical variables: Tense choice in English. In R. Steele & T. Threadgold (Eds.), *Language Topics: Essays in honour of Michael Halliday* (pp. 281–306). John Benjami. https://doi.org/10.1075/z.lt2

Rose, D. (2014). Analysing pedagogic discourse: an approach from genre and register. *Functional Linguistics*, *1*(1). https://doi.org/10.1186/s40554-014-0011-4

Rose, D. (2019). *Reading to learn: accelerating learning and closing the gap*. Reading to Learn.

Rose, D., & Martin, J. R. (2012). *Learning to write, reading to learn: genre, knowledge and pedagogy in the Sydney school*. Equinox Publishing Ltd.
Scott, M. (1997). PC analysis of key words – and key key words. *System*, *25*(2), 233–245. https://doi.org/10.1016/S0346-251X(97)00011-0

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins. https://doi.org/10.1075/scl.22

Zappavigna, M., Martin, J. R., & Springerlink (Online Service. (2018). *Discourse and Diversionary Justice: An Analysis of Youth Justice Conferencing*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-63763-1

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.