LEITURAS TEXTOMÉTRICAS EM SEMÂNTICA INTERPRETATIVAPRINCÍPIOS E PERSPECTIVAS DO PROJETO *SITTELLE*

Por Mª de Fátima B. de M. Batista e Raquel Barbosa de Mesquita Batista,
tradução para o português do original francês de PINCEMIN,B In ASEL/UFPB, vol.27,nº 2,ano 46; 2022º

1. Desafio científico

1.1. Intuição inicial e motivações

Não é fácil para um leitor iniciante, ou mesmo para um linguista familiarizado com a semântica interpretativa, orientar-se no grande corpo dos escritos de François Rastier, ou pensar que compreende, suficientemente, em todas as suas ressonâncias, tal termo, conceito ou passagem. Assim nasceu a ideia do projeto Sittelle, sigla para "Semântica Interpretativa e Textual": experimentar, da forma mais aberta possível, como um corpus numérico, dotado de ferramentas textométricas, pode dar acesso a percursos metódicos e apropriados a textos de Rastier, como aliás, ele se compromete, a respeito de outros autores, em seu livro sobre a semântica do corpus (RASTIER, 2011).Nossa hipótese é que tal ambiente poderia atender às necessidades da pesquisa científica (aprofundamento de conceitos, busca de uma passagem precisa ou de uma citação apropriada, estudo sistemático de passagens) e constituir uma fonte pedagogica (esclarecimento de uma passagem por outra, pesquisa de formulações canônicas como definições, (repérage) identificação de variantes mais desenvolvidas ou simplificadas, ou em um contexto mais claro, ou com exemplos mais expressivos, etc.),(amont[1] ou en aval[2] up stream ou a jusante das sínteses didáticas.

Assim, por ocasião do 35º aniversário do livro Sémantique interpretativa (RASTIER, 1987) e com base em nossa própria pesquisa, pareceu-nos mais estimulante não centrar nossas observações em elementos de leitura pessoal da semântica interpretativa, mas sim compartilhar novos meios de leitura e análise que abrem múltiplas possibilidades para que cada um trace seu próprio caminho na obra de François Rastier, de acordo com suas expectativas e necessidades.Não se trata apenas de trabalhar na criação de um corpus instrumentalizado, mas também, de refletir, em conjunto, sobre os tipos de interrogação e cálculos analíticos relevantes sobre tal corpus. E sobre esses dois aspectos (criação e exploração do corpus), um empreendimento colaborativo e

1. Um corpus numérico trabalhado poderia ajudar na recolha sistemática de materiaais, na sua organização, séleção, qualificação (por seu contexto local — na passagem — et global — na obra de Rastier —, sua frequência, etc.). Pensemos, por exemplo, no projeto de envergadura o *Dictionnaire sémiotique en ligne* (**https://semiotique.org**).

2. Uma apresentação redigida e organizada como (HÉBERT, 2001) ; ver, também, as introduções à semântica interpretativa (publicadas com a rubrica *Repères pour l'étude du site Texto!*) é un ponto de entrada privilegiado que permite colocar os grandes pontos de referência e de introduzir os conceitos-chave ; o prolongamento pode ser, em seguida, um reencontro direto com os próprios escritos de Rastier, um retorno à sua leitura, tanto tradicional, quanto assistida por ferramentas de pesquisa e de exploração.

organizado, que coloca em destaque edições de texto homogêneas e que agregam referenciais metodológicos e interpretativos parece ser o caminho mais gratificante.

Talvez seja útil especificar, em princípio, o que distingue a textometria dentro de abordagens instrumentalizadas para textos numéricos, para entender que aspectos lhe parecem particularmente apropriados. A textometria oferece visualizações sintéticas, que não pretendem substituir a leitura do texto, mas sim revivê-lo, segundo uma dialética que caminha entre quantitativo e qualitativo, resultados de cálculos e retorno ao texto, *leitura distante e leitura fechada* (HEIDEN, 2004). Ele fornece os meios para pesquisar, sistematicamente, as ocorrências de uma palavra ou de uma formulação e, assim, confirmar, fatualmente, as hipóteses que foram explicitadas: trata-se de uma abordagem que já apreciamos com as edições numéricas, da qual a textometria pode dar uma versão avançada (corpus unificado que leva em conta intertextualidade vs buscas múltiplas em textos ou capítulos isolados; busca por padrões complexos que consideram informações morfossintáticas ou estruturais dos textos versus busca por cadeias de caracteres).A textometria também oferece ferramentas heurísticas (formas de apresentação de texto, ordenação) e cálculos estatísticos, para destacar regularidades ou singularidades despercebidas (MAYAFFRE, 2007b) e gerar novos observáveis (RASTIER, 2011, 2020). Por fim, não visa automatizar um modelo de análise de texto com objetivo de precisão e desempenho, pois não prevê uma única leitura com a extração do "sentido", mas confia a condução da análise e a elaboração da interpretação para o pesquisador. Em tudo isso, portanto, ela se destaca do *textmining,* edições digitais padrão, como o tratamento automático das linguas ou a *web semântica* (BÉNEL, 2017), com posicionamento e objetivos diferenciados, geralmente mais próximos de leituras científicas ou atividades educativas. Atenta ao texto e aberta ao trabalho interpretativo, participa de uma filologia numérica e de uma hermenêutica numérica (MAYAFFRE, 2007a).

1.2. Características relevantes do corpus de escritos linguísticos de François Rastier

Certo número de características do trabalho científico de François Rastier encorajam tal abordagem

Uma primeira característica bastante óbvia é a sua volumetria: a obra é prolixa, mas se pode ter um conhecimento global dela. A totalidade dos escritos científicos de François Rastier, que começou a publicar no final da década de 1960 e não parou de escrever nos últimos cinquenta anos, representa, atualmente, cerca de seiscentos artigos ou capítulos de livros e dezessete livros.No entanto, esses escritos exploram questões diferentes que desenvolvem, às vezes, uma relativa autonomia dentro de um pensamento unificado: questões reativadas pela dinâmica da pesquisa contemporânea - **estudos semióticos**, **ciências cognitivas,** linguística dos **corpora numéricos**, cruzando com um entusiasmo geral por **ontologias**, em relação às quais. as proposições da **semântica interpretativa** podem ser elaboradas; a relativização epistemológica das fronteiras disciplinares e um programa unificador das **ciências da cultura**, que interessa, tanto à pesquisa quanto à educação; a releitura de **Saussure** após a descoberta de seus manuscritos e seus esclarecimentos do cèlebrte *Curso de Lingüística Geral*, editado por seus alunos; as questões da escrita e as relações entre as obras —**criação, tradução, transmissão.**..— um método de abordagem linguística de **textos literários,** com atenção especial à escrita de testemunhas e sobreviventes do extermínio (em particular **Primo Levi**) e, em contraponto à

análise da produção dos pensadores, nos quais se pode evidenciar a ligação com essas ideologias destrutivas (**Heidegger**), a caracterização da expressão da xenofobia na Web (*projet princip .net* para filtrar sites racistas) e uma denúncia crítica dos movimentos contemporâneos que atravessam o mundo científico e a sociedade — **desconstrução, decolonialismo, estudos de gênero, cultura do cancelamento.**Como o próprio nome sugere, o projeto Sittelle gostaria de se concentrar na semântica interpretativa e na semântica dos textos en *corpus* (sem excluir a perspectiva de abertura para outros autores), e não no pensamento global de François Rastier. O corpus previsto para Sittelle, portanto, não visa a toda a bibliografia de Rastier, mas a coleção dos escritos mais linguísticos. É claro que a delimitação do corpus procederá ao ajuste do limiar entre o "suficiente" e o "insuficiente" linguístico, se pensarmos na construção do corpus como uma dinâmica, sobretudo das prioridades sobre os textos a serem integrados. Se fizermos uma estimativa, com base em obras, um primeiro círculo de prioridade poderia dizer respeito à *Semântica Interpretativa* (1987) (1994), 2001), *A medição e o grão* (2011). Mas podemos encontrar desdobramentos complementares na perspectiva literária de *Sentido e textualidade* (1989) e *Mundos às avessas* (2018), na discussão da semântica ontológica em (1991) e *Fazer sentido* (2018), nos cruzamentos com o pensamento saussureano (*Saussure au futur,* 2015), nos estudos semióticos preliminares de *Idéologie et théorie des signes* (1971) e *Ensaios de semiótica discursiva* (1973/1974), etc. À primeira vista, o corpus Sittelle poderia, portanto, pretender reunir cerca de uma centena de textos (do tipo artigo ou capítulo). E para percorrer, metodicamente, por uma centena de textos (ou mesmo algumas centenas, se necessário), a ferramenta textométrica parece bem-vinda.

Uma segunda característica relevante do corpus Rastier é sua unidade de escrita e de pensamento, que confere coerência e valor interno ao todo. O que dificulta a delimitação do corpus, também o torna interessante: ressonâncias, repetições, afinidades nos conceitos mobilizados, o modo de pensar, o de se expressar, através das diferentes questões abordadas, os diferentes campos estudados. Isso nos coloca em condições favoráveis para observações, em parte baseadas em palavras: o significado de uma palavra pode, é claro, variar de acordo com seus contextos (aferência), mas, exceto em casos de homonímia, podemos levantar a hipótese de uma certa estabilidade ou continuidade de sentido entre as ocorrências (inerência) e a pertinência em reunir seus contextos para enriquecer sua descrição semântica. Este é o princípio hermenêutico das "passagens paralelas", aplicáveis dentro de um texto ou de uma intertextualidade construída e interpretada.

Uma terceira característica a ser mencionada talvez seja menos óbvia, mas não menos decisiva para o nosso projeto. É que, com a obra de François Rastier, nós trabalhamos um estilo de escrita que funciona mais por evocação, comparações, repetições, retomadas do que por organização progressiva e hierárquica, formal e explícita. A liberdade de pensamento do autor flui de uma declaração flexível e estruturada sem ser conduzida por uma estrutura pré-estabelecida. O próprio autor descreveu as atividades de expressão e de interpretação como mecanismos de ajuste, correção, adaptação (RASTIER, 2001, p.49-50; RASTIER, 2011, p.61). Temos expectativas (expressivas ou interpretativas) e o sentido é construído numa reelaboração, num ajuste progressivo do conteúdo. Inversamente, seria uma concepção ontológica da semântica, contrária à praxeologia afirmada pela semântica interpretativa, que nos permitiria pensar que uma "coisa" (ideia, conceito) pudesse ser, inteiramente, expressa ou compreendida em uma única enunciação e que o movimento avançar em um texto seria uma acumulação ordenada de conteúdo. A renovação do pensamento e o enriquecimento do sentido ocorrem na encruzilhada

de uma expressão múltipla – nem única e definitiva, nem estritamente posicionada e parada em um sistema estático que bastaria desdobrar. E é, justamente por isso, que a leitura do nosso corpus não se contenta com um percurso linear, mas é enriquecida por repetições intertextuais, que as navegações transversais permitidas pela ferramenta textométrica nos parecem ser uma grande contribuição para ampliar e esclarecer o trabalho hermenêutico que o leitor já realiza com sua memória. As formulações locais, ou a entrada no universo da teoria semântica de Rastier, recebem esclarecimentos para referências (réperes) globais.

1.3. Notas

Já se pode perceber que a abordagem textométrica faz operar uma forma de colocação em abismo da semântica interpretativa: que o global determina o local é um princípio da semântica interpretativa, mas, também aqui, torna-se um modo instrumentalizado de acesso à interpretação do corpus dos escritos de Rastier. De fato, as afinidades da teoria da semântica interpretativa de Rastier, por um lado, e da metodologia textométrica, por outro, são múltiplas e importantes (PINCEMIN, 2010).

Tal iniciativa é igualmente possibilitada pela generosidade e abertura de espírito do nosso autor, que não só aceita acompanhar-nos e apoiar-nos neste trabalho (acesso a documentos digitais de origem, partilha de direitos, aconselhamento sobre a composição do corpus, feedback nas observações), mas também consente em expor seus escritos aos mecanismos brutos e potencialmente brutais do processamento formal, não deixando escapar nenhum detalhe da materialidade linguística dos textos, o que pressupõe usos esclarecidos e benevolentes, que devemos incentivar, mas não podemos garantir . Este tipo de publicação é profundamente altruísta, beneficiando sobretudo estudantes e investigadores

2. Perspectiva metodológica

Concretamente, que novos modos de navegação podem ser vislumbrados com um ambiente textométrico do corpus? Esboçamos uma primeira gama de tipos de questionamento, com base num corpus ilustrativo denominado “Zero”, rapidamente reunido numa base exploratória para um primeiro feedback experimental. Todo o processamento é realizado com dois softwares de código aberto: o software TXM 0.8.1 (seções 2.2 a 2.6) (HEIDEN et al., 2010), e o software IRaMuTeQ 0.7 alpha 3 (seção 2.7, sobre o método Reinert) (RATINEAU & DEJEAN, 2009; RATINEAU, 2018).

2.1. Corpus Zero Ilustrativo

O corpus Zero é composto por textos postados no site Texto!(4), difundidos em formato HTML(5)e que desenvolve considerações linguísticas. Reúne vinte e oito textos e apresenta um tamanho de 320.000 palavras (para dimensioná-lo de acordo com as unidades usuais da linguística do corpus).

O detalhe da composição do corpus e as operações de preparação para edição textométrica são dados no anexo 1. Para esta primeira visão geral no início do projeto, usamos uma representação bastante resumida e simplificada dos textos, em formato de texto simples. TXT). A edição XML TEI levará mais tempo para ser alcançada, mas expandirá e refinará as possibilidades de processamento

textométrico. Inicialmente, a experimentação em corpus de texto bruto já pode lançar luz sobre elementos que serão importantes a serem levados em conta nas escolhas de estruturação da TEI.

Durante a importação para TXM, uma etiquetagem morfossintática usada pelo software TreeTagger (SCHMID, 1994) é realizada em tempo real. Ele associa cada palavra ao seu lema (sua forma não flexionada encontrada no verbete de um dicionário) e sua categoria gramatical, de acordo com o conjunto de etiquetas do modelo geral francês, construído por Achim Stein, distribuído no site do software. Estas informações linguísticas podem ser utilizadas na formulação de buscas, como nossos exemplos irão mostrar.

2.2. Concordância

A concordância coleta e apresenta, de forma sintética, todos os contextos de uso de uma determinada palavra (ou qualquer padrão de busca). Em uma ferramenta de textometria como o TXM, a exibição dos resultados permite visualizar, efetivamente, as repetições e variantes, graças a uma apresentação de tabelas que alinha e sobrepõe os contextos (figura 1)e às múltiplas possibilidades de ordenação adaptadas à natureza linguística dos os dados. Com este dispositivo de exibição heurística, a exploração de contextos vai muito além das possibilidades de navegação oferecidas pelos motores de busca das ferramentas buróticas generalistas, sem contar também no interesse de realizar pesquisas que percorrem todo o intertexto, não se limitando a cada arquivo de texto.

As concordâncias nos parecem uma implementação particularmente eficaz da técnica hermenêutica das passagens paralelas, tão esclarecedora para estudar o significado de uma palavra ou de uma formulação a partir de um estudo sistemático de seus usos (PINCEMIN, 2007a). São simples de operar e oferecem um novo ponto de entrada na obra — por palavra ou motivo —, particularmente interessantes também do ponto de vista didático.

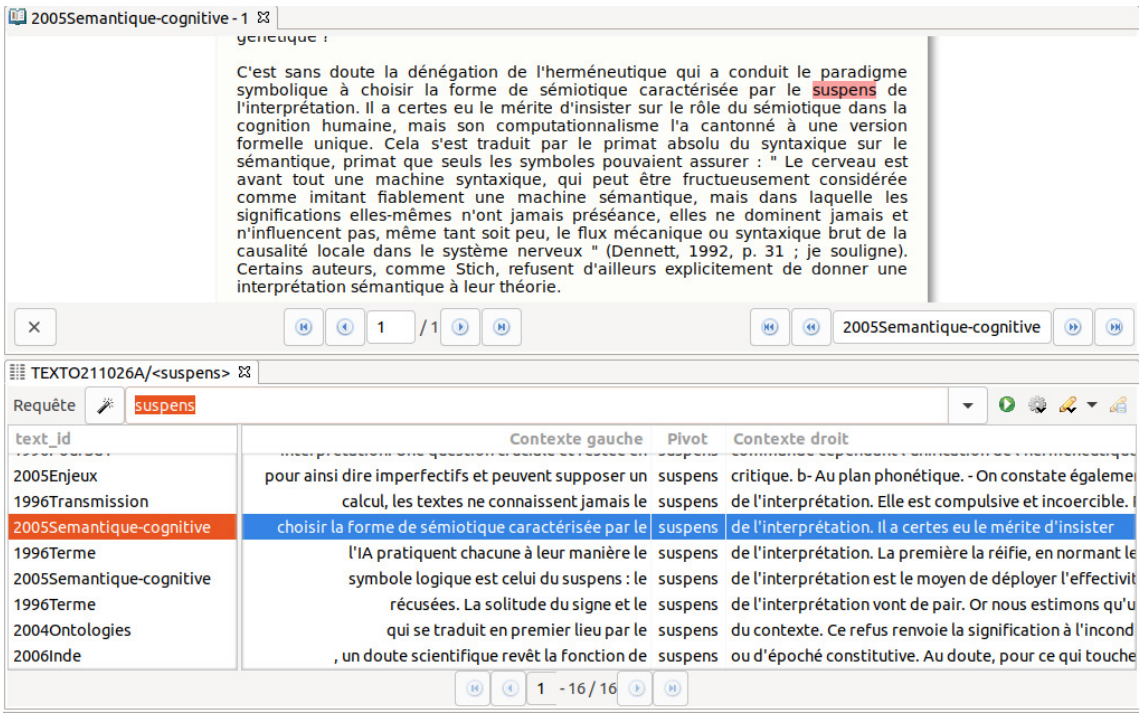

Figura 1: Concordância da palavra “suspenso”, ordenada no contexto correto.

Um primeiro exemplo é dado para uma concordância sobre a palavra “suspenso” (figura 1). Todas as ocorrências da palavra são reunidas na tabela de resultados, com seu contexto imediato: ou seja, 16 linhas para 16 empregos da palavra (contagem dada abaixo da tabela). A localização dos trechos correspondentes é dada na coluna da esquerda: como preparamos o corpus, essa localização indica o código do texto em questão (as indicações podem ter outra forma ou ser mais precisas, por exemplo, com indicação de paginação, se os dados são preparados em conformidade).Essas ocorrências são apresentadas inicialmente ao longo dos textos e na ordem de constituição do corpus (aqui a ordem alfabética dos nomes dados aos textos), mas, para um trabalho analítico, pode-se ordenar esses extratos de acordo com as palavras que seguem (ou que precedem) o nosso pivô “suspense”: assim reunimos todas as passagens que evocam o “suspense da interpretação”. Para compreender plenamente esses trechos, podemos precisar de um contexto mais amplo, ao qual temos acesso, clicando duas vezes na linha em questão: a edição completa do texto é exibida, posicionada no nível do trecho, e com o termo procurado em destaque.

O mesmo mecanismo pode ser estendido para pesquisas de padrões mais complexas. Por exemplo, o idioma de pesquisa nos permite solicitar todas as passagens que mencionam “global” e “local” pelo menos de 10 palavras (d`´écart) diferentes.Encontramos 68 delas em nosso corpus Zero e, ordenando no pivô, podemos identificar as principais formulações associadas: “o global determina o local” (6occ. em 6 textos diferentes), “a determinação do local pelo global” (4 occ.), “a determinação/impacto do global no local” (6occ.), “(acesso/relacionamento) do global ao local” (7 occ.), ou seja, quatro fórmulas que somam mais de um terço das menções que associam global e local. O exame dos contextos também mostra, claramente, que essa determinação do local pelo global tem o valor de um “princípio hermenêutico” (1996Temas, 2003-Ontologia-semiótica, 2005Microssemântica, 2005Mesossemântica).

2.3. Visão geral do local da ocorrência

A concordância, ao apontar para cada contexto onde se localiza, já permite perceber, por exemplo, se as ocorrências encontradas estão agrupadas principalmente em um ou dois textos, ou se estão distribuídas ao longo do corpus. Esta afirmação é completa e precisa, mas podemos ter uma restituição adicional e mais visual com funcionalidades dedicadas a esta questão do arranjo das ocorrências.

Tomemos o caso da noção de isossemia. Uma concordância revela 40 ocorrências do termo: 39 no texto sobre mesossemântica, e uma naquele sobre Borges — ocorrência isolada com contexto de lembrete definitório sintético, “as isossemias ou acordos sintáticos em gênero e número”. Em relação às 39 ocorrências de um mesmo texto, uma visualização permite entender que elas estão concentradas em uma seção do texto (figura 2).O desenvolvimento do texto é representado pelo eixo das abcissas (os números correspondem às posições das palavras na sucessão de textos do corpus), e cada “passo” traduz uma ocorrência do termo. Assim, a subida relativamente acentuada no 1º terço do texto abrange desde a secção 3.1 (*Relações de concordância/A construção de fundos semânticos*) até à secção 4.3. (*Exemplos e problemas de descrição*). A seção seguinte (5) é dedicada às formas semânticas, sem mais ocorrência de “**isossemia**” (a curva tem um longo platô plano). A ocorrência isolada no final do texto está, na verdade, em uma nota referida na seção 6.

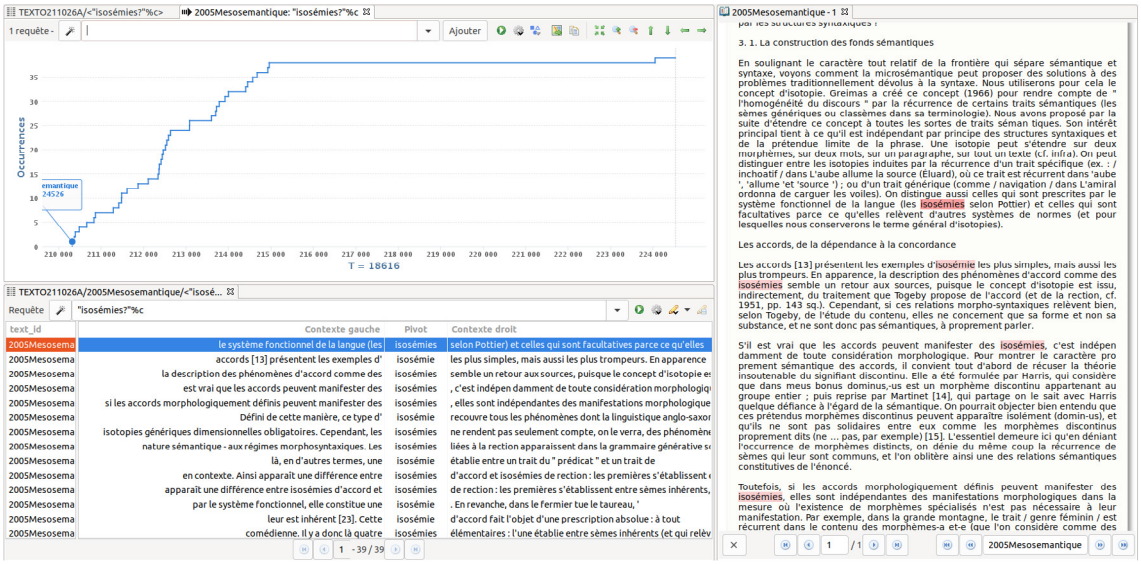

Figura 2: Gráfico de progressão (canto superior esquerdo) das ocorrências de “isossemia(s)” no texto Mesossemântico de 2005.

2.4. Inventários lexicais

As funcionalidades de interrogação textométrica também permitem elaborar inventários gerais **de** vocabulário ou observar como é realizado um determinado padrão.

Para ilustrar o primeiro cenário (inventários gerais), solicitamos os 20 substantivos, adjetivos e verbos mais frequentes em nosso corpus Zero. Isso produz listas de palavras fora de contexto, mas o software se preocupa em fornecer acesso direto (por link de hipertexto) à visualização dos contextos de uso (em concordância). Por exemplo, podemos verificar que a maioria das 655 ocorrências de “gênero(s)” referem-se à noção de gênero textual (com apenas cerca de dez ocorrências da frase “este gênero de» e algumas menções ao gênero gramatical (masculino / feminino)).

Tableau 1 : Les 20 noms communs, adjectifs qualificatifs et verbes les plus fréquents du corpus Zéro (avec indication de leur fréquence).[3]

Noms		Adjectifs	Qualificatifs	Verbes	
Texte	1420	sémantique	794	Etre	5469
exemple	814	même	729	avoir	2118
genre	655	autre	658	pouvoir	1404
sens	596	linguistique	495	faire	572
forme	594	textuel	9	permettre	464
langue	586	cognitif	235	rester	427
discours	524	scientifique	233	définir	421
mot	516	interprétatif	230	devoir	350
langage	459	générique	218	relever	288
sémantique	457	général	214	dire	280
théorie	433	divers	202	décrire	262
relation	412	sémiotique	202	trouver	236
corpus	398	propre	195	considérer	217
unité	388	social	191	distinguer	217
signe	362	nouveau	175	falloir	217
sème	349	herméneutique	171	mettre	211
contexte	344	grand	165	prendre	202
linguistique	342	logique	165	constituer	200
concept	341	différent	164	opposer	189
objet	334	humain	155	représenter	182

3. Nós retificamos as espécies brutas, retirando da lista dos nomes as abreviaturas « cf » et « p », e reescrevendo « sens » o lema etiquetado de forma inutilmente complexa « sen|sens ».

Tabela 1: Os 20 substantivos comuns, adjetivos qualificadores e verbos mais frequentes do Corpus Zero (com indicação de sua frequência).

Substantivos		***Adjetivos***	Qualificativos	Verbos	
Texto	1420	semântica	794	ser estar	5469
exemplo	814	mesmo	729	Ter	2118
amável	655	de outros	658	potência	1404
sentidos	596	linguístico	495	fazer	572
Formato	594	textual	9	permitir	464
Língua	586	cognitivo	235	ficar	427
Fala	524	cientista	233	definir	421
palavra	516	interpretativo	230	Trabalho de casa	350
Língua	459	genérico	218	levantar	288
semântica	457	em geral	214	dizer	280
teoria	433	vários	202	descrever	262
relação	412	semiótico	202	encontrar	236
corpus	398	ter	195	considerar	217
unidade	388	social	191	distinguir	217
sinal	362	novo	175	precisar	217
semear	349	hermenêutica	171	colocar	211
contexto	344	grande	165	levar	202
linguístico	342	lógica	165	constituir	200
conceito	341	diferente	164	opor	189
objeto	334	humano	155	Representar	182

Esse tipo de lista vai confirmar o que o especialista já suspeita (como a dominância de “texto” e do textual”, e de “semântica”; e, quanto aos verbos, a precedência de auxiliares e modais), mas pode também recordar a importância (pelo menos a simples presença quantitativa) de uma noção (se tivéssemos consciência do lugar ocupado pelo “exemplo” na escrita científica de Rastier), ou mesmo pode chamar a atenção para os pontos de entrada que não se esperava estar em um lugar tão bom (aqui talvez as considerações “sociais” que, de fato, combinam com as “ciências sociais” (42 occ.), a “prática(s) social(s)” (66), a “norma social(s)” (10), a “demanda social” (8), que se estendem, também ,a quase um quarentena de outros nomes).

Pode-se, de fato, elaborar listas de todos os tipos de padrões linguísticos, mais ou menos complexos e mais ou menos direcionados. Cabe ao pesquisador definir qual “rede” ele quer lançar sobre o corpus. Aqui, por exemplo, podemos estar interessados em inventariar expressões frequentes

que podem formar um termo (com um padrão gramatical do tipo Substantivo+Adjetivo), ou mesmo listar grupos nominais formados com o núcleo substantivo "linguístico" ou "semântico".(Mesa 2).

Tabela 2: Extratos das realizações de padrões definidos por diferentes informações linguísticas (morfológicas, lexicais, sintagmáticas).

Séquences Nom+Adj	(fréq. ≥ 2_)	Sémantique/linguistique +	***Adj./ Cpl. de nom***
parcours interprétatif	97	sémantique cognitive	78
forme sémantique	96	sémantique des textes	51
molécule sémique	89	linguistique de corpus	45
sémantique cognitive	81	sémantique interprétative	20
champ générique	79	linguistique historique	17
pratique sociale	66	linguistique générale	15
isotopie générique	51	sémantique linguistique	14
texte scientifique	45	sémantique structurale	14
positivisme logique	44	sémantique lexicale	13
science sociale	42	linguistique cognitive	10
trait sémantique	38	sémantique formelle	10
impression référentielle	37	linguistique textuelle	9
fond sémantique	34	sémantique différentielle	9
nom propre	34	linguistique du signe	7
relation sémantique	32	linguistique de la parole	6
traitement automatique	32	linguistique des textes	6
discours scientifique	31	sémantique diachronique	6
sème générique	31	sémantique textuelle	6
sème inhérent	31	linguistique du texte	5
sens littéral	31	linguistique saussurienne	5
unité sémantique	31	sémantique procédurale	5
roman policier	30		
classe lexicale	29	***Mots commençant*** par « isotop »	
domaine sémantique	29	isotopie(s)	241
forme textuelle	29	isotope(s)	4
sème afférent	29	isotopique(s)	3
perception sémantique	28		
problématique logico-grammaticale	28		

Sequencias Nome+Adj	(fréq. ≥ 2_)
percurso interpretativa	97
forma semântica	96
molécula sêmica	89
semântica cognitiva	81
campo genérico	79
prática social	66
isotopia genérica	51
texto científico	45
positivismo lógico	44
Ciências Sociais	42
recurso semântico	38
impressão referencial	37
fundo semântico	34
nome próprio	34
relação semântica	32
processamento automático	32
discurso científico	31
sema genérico	31
semente inerente	31
significado literal	31
unidade semântica	31
história de detetive	30
classe lexical	29
domínio semântico	29
formulário de texto	29
sema aferente	29
percepção semântica	28
problema lógico-gramatical	28

Semantica/Linguistica +	Adj./ Cpl. de nome
semântica cognitiva	78
semântica de texto	51
linguística de corpus	45
semântica interpretativa	20
linguística histórica	17
linguística geral	15
semântica linguística	14
semântica estrutural	14
semântica lexical	13
linguística cognitiva	10
semântica formal	10
linguística textual	9
semântica diferencial	9
linguística de sinais	7
linguística da fala	6
linguística de textos	6
semântica diacrônica	6
semântica textual	6
linguística do texto	5
linguística saussuriana	5
semântica processual	5
Palavras começando por “isotop”	
isotopia(s)	241
isótopo(s)	4
isotópico(s)	3

Podemos também trabalhar, no plano morfológico, e explorar as variantes derivacionais de« isotopia »[4] atestadas no corpus, para atrair observações, tais como : o uso dominante diz respeito ao nome (« isotopia(s) ») ; aliás, tem-se raros adjetivos « isotópico(s) », que significam em relação de isotopia (« ser isotópico », « os contextos isotópicos de um semema »), e « isotópícos(s) », relativo (s) à isotopia (« análise/tipologia/feixe isotópico »), mas nenhum « isotopant(e)(s) , por exemplo, em nosso corpus Zéro.

4. Conta CQL utilizada : «`isotop.*»%c`

2.5. Coocorrências

O cálculo de coocorrências pode ser útil para operar uma síntese estatística dos contextos de uma palavra (ou de uma expressão, de um tema), principalmente quando ela é muito frequente e pode ajudar a ter uma visão geral. Se, por exemplo, eu pedir as co-ocorrências de isossemia(s), a lista de resultados do cálculo (figura 3) me lembrará um certo número de noções associadas: isotopia, concordância, reação, facultativo/obrigatório, aferente/ inerente, etc

Vamos ilustrar, ainda, essa funcionalidade com a busca das palavras coocorrentes da família "interpretação" ("interpretação", "interpretativo", "interpretar", etc.) (figura 3). Os resultados chamam-me a atenção para a colocação em correspondência da interpretação com a produção (declinadas nas diferentes categorias gramaticais: "produção", "produtivo", "produzir"), e olhando para os contextos podemos confirmar que a proposta associa cerca de trinta vezes " a produção ou/e a interpretação" dos textos.

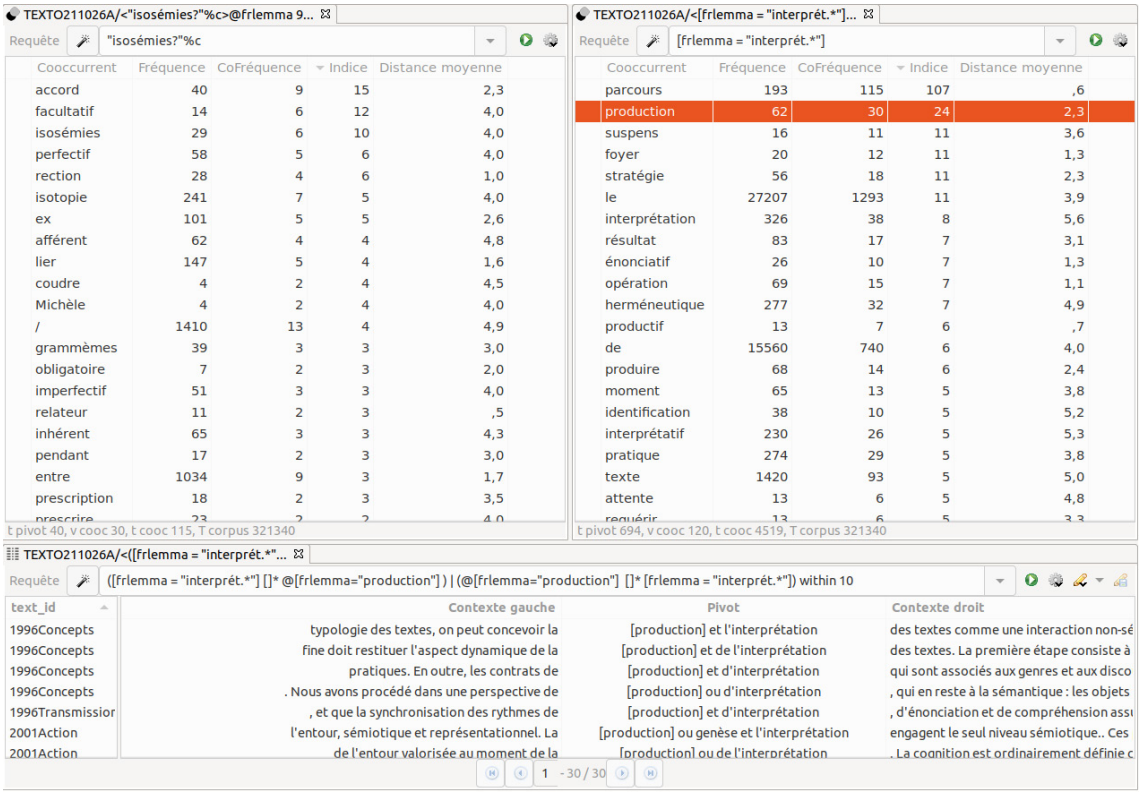

Figure 3 : Coocorrentes de "isossemia" (à esquerda) ou palavras da família de "interpretação" (à direita). Na parte inferior, a apresentação em concordância os contextos de "produção" com uma palavra da família de "interpretação", calculada automaticamente ,clicando duas vezes na linha do co-ocorrente ("produção").

2.6. Mapeamento e caracterização intertextual

Ao lado das sínteses estatísticas, as análises fatoriais são capazes de produzir visões de conjunto de uma coleção de textos, na forma de mapas, destacando os principais contrastes que estruturam a coleção em questão. Dependendo do conjunto de textos considerado, poder-

se-iam trazer à tona desdobramentos mais cronológicos ou mais temáticos, por exemplo, ou ainda oposições de gêneros textuais. Além disso, o cálculo das especificidades fornece uma caracterização lexical de cada texto, útil, por exemplo, para orientar as escolhas de leitura: "este texto parece abordar particularmente essas noções".

Essas cálculos de análises de dados e de estatística, que fornecem uma visão de conjunto do corpus por inteiro, são complementares da leitura tradicional de estudo como destaca Étienne Brunet por ocasião de um trabalho sobre o tempo em um corpus de literatura francesa:

> […] l'ordinateur échappe aux pesanteurs du temps qui entravent la mémoire humaine et qui imposent aux souvenirs une perspective fuyante, où les éléments les plus disponibles sont ceux que fournit l'environnement présent. Là où le lecteur parcourt en marchant l'espace littéraire, dans la succession changeante et l'effacement progressif des paysages, l'ordinateur saisit d'un coup le même espace, comme on lit une carte géographique ou stratégique, tous les points étant à plat, offerts à l'œil en même temps. En réalité les textes, si écrasés qu'ils soient par la perspective plongeante d'un observateur posté sur Sirius, acquièrent une lisibilité qu'ils n'ont pas pour l'explorateur engagé dans le maquis de la lecture. Au ras du sol, en enjambant les ruisseaux, on peut difficilement délimiter la ligne de partage des eaux. Mais d'en haut le paysage littéraire se découvre avec l'orientation des chaînes, les pentes, les ruptures et tous les mouvements de terrain produits par l'histoire. (BRUNET, 2016, p. 371)[5]

[...] o computador escapa aos pesos do tempo que entravam a memória humana e que impõem às memórias uma perspectiva fugaz, onde os elementos mais disponíveis são os fornecidos pelo ambiente presente.Ali onde o leitor percorre, atravessando o espaço literário, na sucessão cambiante e no apagamento progressivo das paisagens, o computador, subitamente, apodera-se do mesmo espaço, como se lêem, num mapa geográfico ou estratégico, todos os pontos que são planos, oferecidos aos olhos ao mesmo tempo. Na realidade, os textos, por mais esmagados que possam ser pela perspectiva merguladora de um observador colocado sobre *Sirius*, adquirem uma legibilidade que não têm para o explorador engajado no maquis da leitura. Ao nível do solo, atravessando os córregos, difícilmente, pode-se delimitar a linha de divisão das águas. Mas de cima da paisagem literária, descobrem-se, com a orientação das correntes, as encostas, as rupturas e todos os movimentos do solo produzidos pela história. : Essas cálculos de análises de dados e estatística, que fornecem uma visão do corpus por inteiro, são complementares da leitura tradicional de estudo*ro, caractérisés par 96 noms ou adjectifs les plus fréquents.*[6]

Essas cálculos de análises de dados e estatística, que fornecem uma visão do corpus por inteiro, são complementares da leitura tradicional de estudo

5. O artigo de onde esta citação foi extraída também é um capítulo digital de (BRUNET, 2016) disponível no Texto! (http://www.revue-texto.net/index.php?id=3756) e no arquivo HAL (https://halshs.archives-ouvertes.fr/halshs-01275527v1L'article dont est tiré cette citation est également un chapitre numérique de (BRUNET, 2016) disponible sur *Texto!* (**http://www.revue-texto.net/index.php?id=3756**)et sur l'archive HAL (**https://halshs.archives-ouvertes.fr/halshs-01275527v1**).

6. A representação também foi clareada e esclarecida, filtrando-se os pontos menos úteis (palavra ou texto): por um lado mal representado (cos2< 0,3) nesta projeção plana, otimizado mas reduzido (informação de variação mantida = 17 ,29 + 13,73 = 31,02%), e também por outro lado pouco envolvido nas grandes variações descritas nestas duas primeiras dimensões (ctrb1 < 1% e ctrb2 < 1%).

> sas cálculos de análises de dados e estatística, que fornecem uma visão do corpus por inteiro, são complementares da leitura tradicional de estudo com a orientação das correntes, as pontes, as rupturas e todos os movimentos do terreno produzidos pela história. (BRUNET, 2016, p. 371)

Essa funcionalidade será interessante aqui para nos permitir apreciar a composição do nosso corpus e seus balanços internos. Para a construção deste representação, escolhemos representar os textos pelo perfil de uso de uma centena de substantivos e adjetivos mais frequentes[7]. (figura 4)

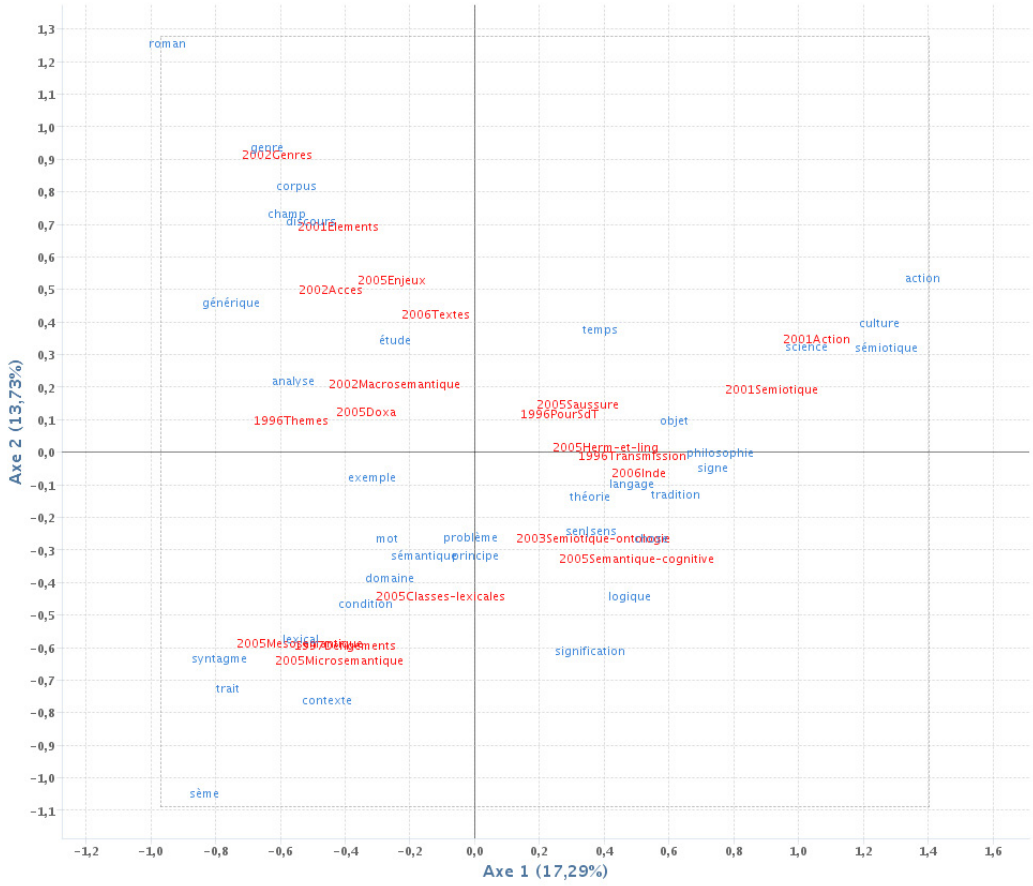

Figure 4 :: Mapeamento por análise fatorial das correspondências dos 28 textos do Corpus Zero, caracterizados pelos 96 substantivos ou adjetivos mais frequentes.

Nosso Corpus Zero parece ser desdobrado em três direções principais: as descrições lexicais da semântica interpretativa (com semas, classes, etc.) desenvolvimentos do lado da semiótica e reflexões sobre o signo, particularmente em relação às ciências cognitivas, mas também aos estudos saussurianos.

Digamos que eu queira identificar o vocabulário que caracteriza um texto de forma diferencial, por contraste com todo o corpus: o cálculo das especificidades identificará as palavras

7. A representação também foi diminuída e esclarecida, filtrando os pontos menos úteis (palavra ou texto): por um lado mal representado (cos2< 0,3) nesta projeção plana, otimizado, mas reduzido (informação de variação mantida = 17 ,29 + 13,73 = 31,02%), e também, por outro lado pouco, envolvido nas grandes variações descritas nestas duas primeiras dimensões (ctrb1 < 1% e ctrb2 < 1%).

particularmente sobre-representadas no texto escolhido (ou mais geralmente em qualquer parte definida no corpus: um subconjunto de textos, um tipo de partes, etc.).

Consideremos, por exemplo, o texto Transmissão de 1996 (Comunicação ou transmissão?), a Tabela 3, logo abaixo apresenta as palavras específicas deste texto com pontuação estatística de 8 ou mais. Podemos observar que o cálculo estatístico relativiza a alta proporção de ocorrências de uma palavra em um texto ao considerar também sua frequência: por exemplo, considera o uso exclusivo de "tradução" (10 occ.) as 108 ocorrências de "comunicação", que teriam sido muito menos possíveis ao acaso. Permite, assim, chamar a atenção para casos que são estatisticamente notáveis, mas que poderiam passar despercebidos a olho nu, ou com uma simples regra de três.

Tableau 3 :Especificidades do texto da Transmissão de 1996 no corpus Zero, considerando todas as palavras do corpus.

Lemme	***Fréquence totale dans le corpus Zéro***	***Fréquence dans le texte 1996 Transmission***	***Indice de spécificité***
communication	108	71	70,6
Transmission	53	33	32,1
Commentaire	38	27	28,8
Traduction	63	33	28,8
Message	28	23	27,2
Code	38	23	22,1
Transcodage	18	16	20,2
Information	59		19,0
Modèle	137	33	16,1
Translation	10	10	13,9
Transmettre	22	12	11,1
Interprétation	326	43	10,9
Le	27207	1308	10,6
émetteur[8]	20	11	10,3
Métalangage	27	12	9,7
Tradition	289	35	8,0

8. A saída bruta do software sugeriu "emissor". Para a legibilidade dos resultados, substituímos por "emissor", definido como um lema que reune as 6 ocorrências de "Emissor" e as 14 de "emissor" (a menos que tenha sido realizado processamento automático), e atualizamos as frequências e pontuação estatística na tabela.

Lema	*Frequência Total no Corpus Zero*	*Frequência no texto 1996 Transmissão*	*Índice de especificidade*
Comunicação	108	71	70,6
Transmissão	53	33	32,1
Observação	38	27	28,8
Tradução	63	33	28,8
Mensagem	28	23	27,2
Codificação	38	23	22,1
trancodificação	18	16	20,2
Informação	59		19,0
Modèlo	137	33	16,1
Tradução	10	10	13,9
Transmitir	22	12	11,1
Interprétação	326	43	10,9
A	27207	1308	10,6
transmissor[9]	20	11	10,3
métalinguagem	27	12	9,7
Tradição	289	35	8,0

Les spécificités peuvent également être lues du point de vue des mots plutôt que de celui des textes : autrement dit, pour un mot donné, quel est son profil quantitatif (statistique) d'usage par rapport à l'ensemble des textes ? Considérons par exemple la présence plus ou moins forte (et statistiquement remarquable) des trois mots suivants au sein des 28 textes du corpus Zéro : « corpus », « données », « traitements »[10]. Un diagramme de spécificités nous aide à repérer les quelques textes qui semblent développer et allier plusieurs de ces notions : « corpus » et « données » dans *1996Themes*, *2002Genres*, *2005Enjeux* ; « corpus » et « traitements » dans *2002Acces*.

As especificidades igualmente ser lidas do ponto de vista das palavras mais do que do ponto de vista dos textos: ou seja, para uma determinada palavra, qual é o seu perfil quantitativo (estatístico) de uso em relação a todos os textos? Considere, por exemplo, a presença mais ou menos forte (e estatisticamente notável) das três palavras a seguir nos 28 textos do Corpus Zero: "corpus", "data", "processing". Um diagrama de especificidades nos ajuda a identificar os poucos textos que parecem desenvolver e combinar várias dessas noções: "corpus" e "dados" em 1996Temas, 2002Gêneros, 2005Enjeux; "corpus" e "processamento" em 2002Acesso.

9. A saída bruta do software sugeriu "emissor". Para a legibilidade dos resultados, substituímos por "emissor", definido como um lema que reune as 6 ocorrências de "Emissor" e as 14 de "emissor" (a menos que tenha sido realizado processamento automático), e atualizamos as frequências e pontuação estatística na tabela.

10. Essa escolha de palavras é inspirada na caracterização da classe 4, construída para a classificação de Reinert e apresentada na figura 6.

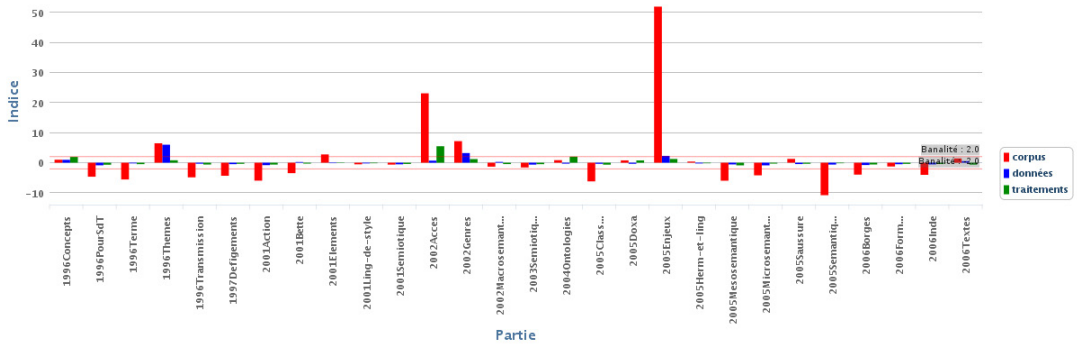

Figure 5 :Diagrama de especificidades das palavras "corpus", "dados" e "processamento" nos 28 textos do Zero corpus.

2.7. Método Reinert e isotopias

Gostaríamos também de experimentar o método de Reinert (REINERT, 1990; RATINAUD, 2018) (algoritmo precursor e próximo aos modelos Topic) para avaliar sua capacidade de esboçar isotopias. De fato, o cálculo define conjuntos de palavras particularmente presentes em fragmentos textuais que tendem a se assemelhar lexicalmente entre si em oposição ao resto dos textos: essas palavras poderiam atualizar semas de isotopias que são particularmente importantes e estruturantes para o conjunto dos textos considerados? As questões metodológicas que surgem prendem-se sobretudo com o ajustamento de parâmetros como o tamanho dos segmentos de texto e o número de classes a solicitar. Lançamos um primeiro experimento adotando a divisão textual padrão (segmentos de cerca de 40 palavras) e perguntando cerca de vinte aulas. O resultado obtido pode ser representado graficamente pela figura 6. Nossa hipótese é que as palavras representativas das classes poderiam sugerir candidatos interessantes para definir isotopias genéricas dominantes dos textos.

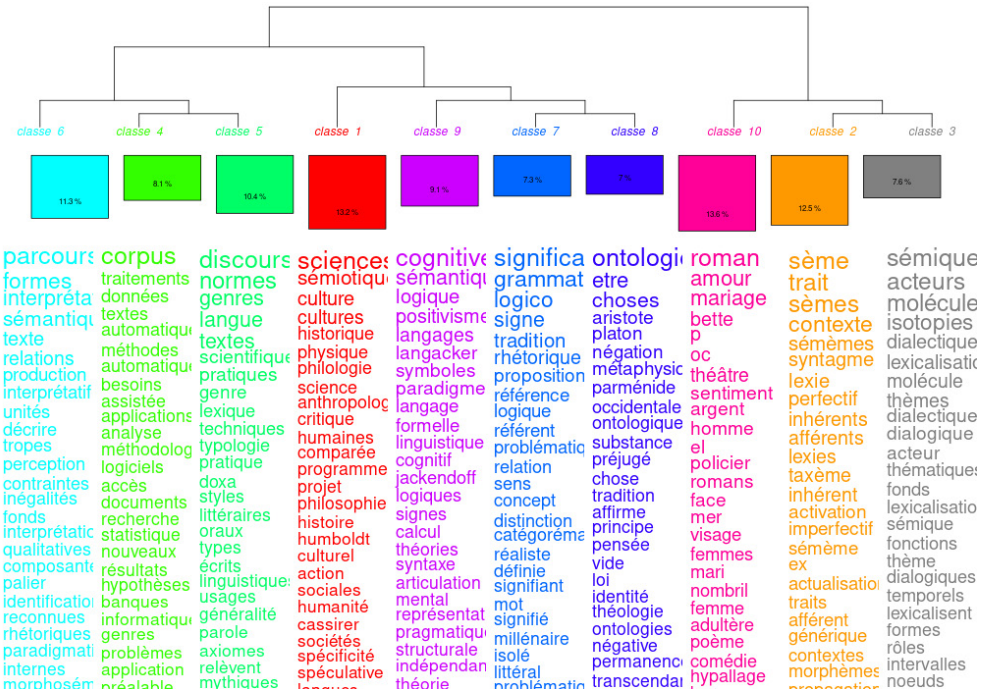

Figure 6 : Classificação de Reinert aplicada ao corpus Zero (com as configurações: segmentos de 40 palavras, 20 classes solicitadas).

Uma trilha complementar e invertida seria uma busca pelas realizações de uma isotopia. Ao invés de formular uma pesquisa a partir de uma palavra ou padrão, a ideia seria trabalhar em encontros de palavras: as palavras se contextualizam e podem ser o traço, o suporte, de isotopias. Propostas nesse sentido foram testadas em pares de palavras (MAYAFFRE, 2008; BRUNET, 2016, capítulo 17). Poder-se-ia considerar conjuntos de palavras (com certa estruturação (PINCEMIN, 1999)) e estudar uma forma de contar com as possibilidades avançadas de um buscador textométrico para identificar manifestações variáveis de um conjunto de lexemas ou morfemas que se contextualizam semanticamente: isso permitiria operacionalizar uma forma de pesquisa menos lexical, mais semântica. Outra pista já pode ser observada com a função "corpus colorido" do IRaMuTeQ (figura 7), que relaciona cada segmento textual a uma classe, potencialmente representativa de sua isotopia dominante.

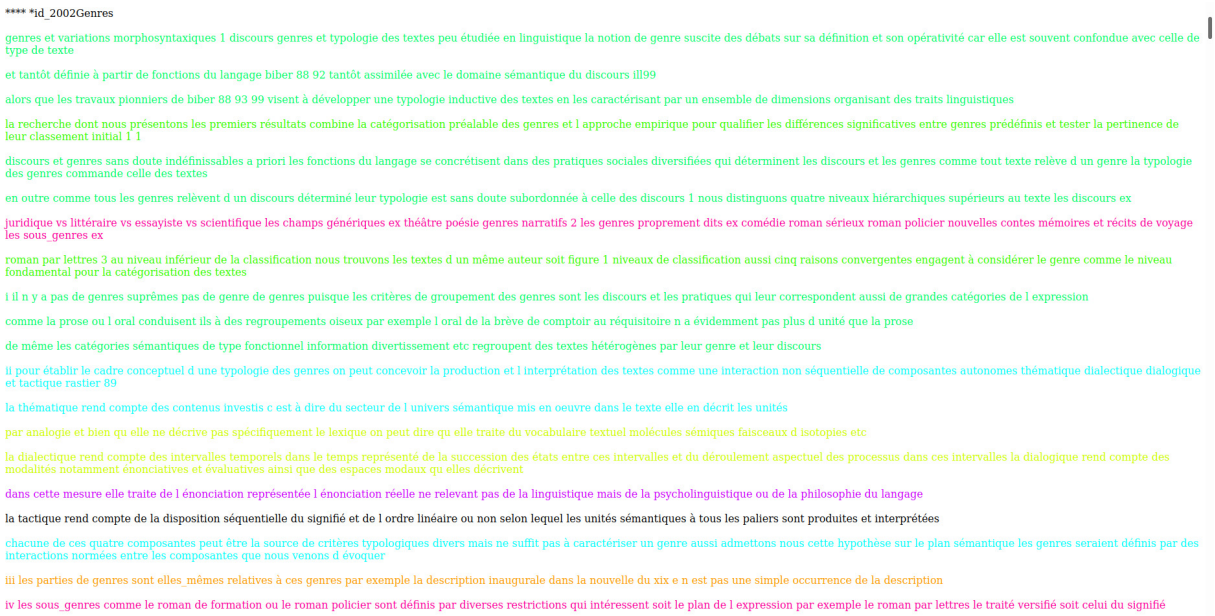

Figura 7 :Função "Corpus em cores" do Iramuteq, resultado para o início do texto Gêneros 2002: aplicação a cada segmento de texto do código de cores da classe à qual foi atribuído pela classificação

3. Princípios orientadores e escolhas técnicas

Essas perspectivas apaixonantes devem ser compartilhadas. A obra de Rastier é de uma riqueza que se presta a múltiplos pontos de vista, e a abordagem textométrica não implementa um cálculo semântico (que identificaria "o" significado do corpus), mas oferece meios de investigação, que cada pesquisador mobiliza de acordo com suas perguntas e hipóteses. Faz sentido investir colaborativamente na produção de um corpus de qualidade, para então poder compartilhar múltiplos caminhos interpretativos construídos de acordo com as diversas expertises de uma grande comunidade de leitores. Trata-se, então, de explicar os princípios orientadores que especificarão nossas escolhas técnicas.

3.1. Progressividade

Começamos com uma pequena seleção de textos, nos quais a abordagem textométrica já pode ser testada; então esse conjunto pode ser ampliado e diversificado gradativamente, possivelmente dando origem a *corpora* e *subcorpora* com geometria variável. Em outras palavras, não se trata de produzir "o" corpus dos escritos de semântica interpretativa de Rastier, mas de alimentar em etapas, ao longo do tempo, uma coleção unificada a partir da qual se construir várias intertextualidades relevantes.

A mesma progressividade é esperada para o modelo textual XML TEI, que se tornará mais preciso com a experiência. Além disso, nem todas as ferramentas textométricas podem ser implantadas desde o início: um primeiro passo será a disponibilização de *corpora* compatíveis com ferramentas que pesquisadores e estudantes possam instalar em seus computadores; e é numa segunda etapa que poderemos considerar o acesso online, dispensando a instalação de softwares e abrindo outras possibilidades, em particular para gerenciar questões de direitos de forma mais fina. O estabelecimento de links com edições online de obras sob controle de acesso nas editoras é uma das questões mais complexas e onerosas *a priori* e pode ser estudada uma vez concluídas as etapas anteriores.

3.2. Texto e intertextualidade

Muito claramente, estamos tomando partido por um corpus de textos, respeitando a unidade contextual do texto integral, em oposição à possibilidade de reunir extratos (se não promovidos à categoria de peças selecionadas) ou mesmo "amostras": a própria semântica interpretativa nos convida a praticar a linguística do corpus como ciência dos textos instrumentados (VALETTE, 2008).

Além disso, na semântica interpretativa como na textometria, o global determina o local. A construção do acervo de textos digitais será, portanto, pautada pela escolha de textos que possam se contextualizar mutuamente de forma relevante. Uma particularidade do nosso corpus é que nos ocupamos com textos vivos: retrabalhados por ocasião de uma nova edição, retomados em publicações abrangentes. Mais do que uma coleção de textos singulares e bem identificados, temos sim uma espécie de rede, de estrutura em árvore, com o desenvolvimento de veios ou galhos. Uma pista seria construir nosso corpus com textos modelares, representativos do desenvolvimento de um sujeito, eventualmente eles próprios redefinidos por ocasião do projeto Sittelle e, portanto, que corresponda mais a um período de escrita-reescrita do que a uma data de publicação. Não parece de todo óbvio ser possível e interessante reunir textos que correspondam, precisamente, a todo um conjunto de publicações bastante completas, ou em todo caso isso constituiria outro tipo de corpus, com outras possibilidades (diacronia mais fina) e outros limites (redundâncias) para as observações.

3.3. Abertura

Do ponto de vista jurídico, o desafio seria tornar acessíveis às consultas Textomêtricas grandes e relevantes conjuntos de textos, compatibilizando-os com os requisitos legais, a fim de disponibilizá-los ao maior número possível de pessoas. Na prática, seria muito mais simples constituir um acervo particular com acesso restrito; mas gostaríamos que o Sittelle fosse justamente a oportunidade de buscar e experimentar soluções para investir, coletivamente, em um corpus

compartilhado, acessível a todos. Estamos, portanto, diretamente interessados em desenvolvimentos recentes da ciência aberta. Alguns textos científicos tornam-se livremente utilizáveis para usos não comerciais. Na França, a *Lei para uma República Digital* deveria facilitar a disponibilização de artigos científicos para ensino e pesquisa.

Para textos protegidos por direitos autorais, poderíamos especificar formas de complementaridade com a edição tradicional para abrir acessos direcionados com boa inteligência, respeitando as restrições econômicas dos editores, de acordo com uma abordagem construtiva e inovadora como por exemplo a OpenEdition[11] et de son programme Freemium[12]*OpenEdition12* e seu programa *Freemium*13 entre os mais avançados. Assim, uma ferramenta de consulta *online* poderia variar o acesso em função dos direitos associados a uma conta de conexão e aos diferentes textos; poderia desdobrar as possibilidades de análise quantitativa e controlar com mais precisão a exibição de contextos, passagens ou páginas. A versão portal do TXM começou a desenvolver tais possibilidades para as necessidades da base do francês medieval , que incluía alguns textos protegidos por direitos autorais.

Em nível técnico, a abertura consiste em optar por formatos digitais padrão, que promove a reutilização dos dados, a interoperabilidade de software e, portanto, a qualidade das trocas científicas (transparência, reprodutibilidade etc.) (GUILLOT et al., 2018) . De fato, seria uma pena produzir dados elaborados e depois impor o acesso mediado por um determinado software, por mais poderoso que fosse; alguns pesquisadores podem querer explorar o corpus por meio de ferramentas complementares e nos parece muito importante permitir isso. O formato TEI destina-se especificamente ao compartilhamento e troca de dados textuais por meio de um padrão internacional. Pode ser explorado, diretamente, por certas ferramentas (como TXM por exemplo); mas para ferramentas baseadas em outros formatos de entrada e para facilitar outros usos, é possível projetar versões do corpus automaticamente derivadas do formato TEI.

3.4. Filologia Digital

A edição digital de textos convida a problematizar a representação do texto e a explicitar as escolhas, pois os "dados" textuais não são dados, "os dados são feitos daquilo que damos a nós mesmos" (RASTIER, 2001, p. 86; ver também RASTIER, 2021): concretamente, de que elementos constitutivos do texto pretendemos dar conta? Assim, os *corpora* se engajam cientificamente em uma filologia digital (RASTIER, 2001; GUILLOT et al., 2017). Para as análises textométricas, além da identificação das unidades linguísticas e textuais necessárias ao processamento (palavras, contextos), questões de leitura e interpretação são centrais. As tecnologias atuais devem, portanto, ser colocadas a serviço de uma restituição das estruturas

11. Open Edition é uma infraestrutura de publicação digital [nacional francesa] que serve à comunicação científica nas ciências humanas e sociais. » Veja a apresentação completa em:: **https://www.openedition.org/6438**

12. Open Edition Fremium é um programa para o desenvolvimento da edição científica de livre acesso no domínio das ciências humanas e sociais. » » Essa empresa que nós propomos, exclusivamente, ás instituições (bibliotecas,campus, centros de pesquisa) visa a conhecer um modelo econômico inovador e durável {...} os textos são acessíveis em livre acesso no formato HTML para todo internauta e são telecarregáveis nos formatos PDF e ePub unicamente não são aplicados

« OpenEditionFreemium est un programme pour le développement de l'édition scientifique en libre accès dans le domaine des sciences humaines et sociales. Ce partenariat, que nous proposons exclusivement aux institutions (bibliothèques, campus, centres de recherche), vise à construire un modèle économique innovant et durable. […] les textes sont accessibles en libre accès au format HTML pour tout internaute, et ils sont téléchargeables aux formats PDF et ePub uniquement ne sont appliqués. » (**https://www.openedition.org/14043**)

textuais essenciais à atividade de leitura e interpretação, e em particular as contextualizações de todos os tipos (intra e intertextual) (PINCEMIN, 2007b), ao invés de empobrecer o texto para "encaixá-lo" nas ferramentas[13].

Este trabalho de edição digital, que exige processamento especializado, assistido, mas não automático, representa, claramente, um certo custo. A textometria precisa desse trabalho? Não funciona tão bem em texto "bruto"? Na verdade, já podemos tirar muitas observações interessantes de um corpus não estruturado como o nosso corpus Zero, como também já encontramos alguns limites. Por exemplo, na figura 1, voltamos ao texto para entender o significado de uma ocorrência de "suspensão". Nesta leitura, temos uma citação, para a qual François Rastier especifica que o sublinhado ou sublinhados são dele. No entanto, a exibição do texto em TXM perdeu esses sublinhados (aqui o itálico). Uma edição digital XML-TEI preservará essas informações e as restaurará para consulta do texto. Da mesma forma, para as partes que eliminamos pura e simplesmente, como a bibliografia: uma edição TEI permite fazê-lo, controlando seu papel na análise textométrica. O caminho de progressão na Figura 2, também, poderia ter sido mais preciso, se pudéssemos codificar as divisões intratextuais: poderíamos então traçar os limites das seções sucessivas no gráfico, ler diretamente na figura o que estávamos procurando no texto. Ainda nos ocorre outro fato que são as correções de erros de digitação: é sempre uma boa ideia simplificar, ignorando-os? Tomemos o caso de uma lista numerada, na qual, encontra-se um erro de numeração, o mesmo número é usado duas vezes: será importante para o leitor saber que, na fonte publicada, o item de número 4 em TXM é na verdade numerado 3 em uma versão lançada publicamente? A edição TEI fornece os meios, se considerados úteis, para acompanhar as correções feitas, sem interferir, de maneira irritante, no processamento da análisc

Um desafio será encontrar o nível certo de modelagem das estruturas textuais. Essa modelagem pretende, simultaneamente, ser leve (por pragmatismo, para que a implementação possa ser realizada em um grande número de textos sem exigir recursos de grande escala) e prudente (para limitar as dificuldades de superinterpretação: quanto mais casos o modelo detalha, mais encontramos a necessidade de determinar, claramente, o caso a ser aplicado em cada situação), além disso produtiva (interessante para análises textométricas).

4. Elementos de conclusão

Como o projeto Sittelle vem a completar, de uma forma original, os modos de disseminação da semântica interpretativa, nosso desejo é que ajude a ampliar e aprofundar o acesso a ela. Poderia permitir, ao mesmo tempo, tanto um estudo dos conceitos e propostas da teoria de Rastier, quanto, talvez até certo ponto, uma experimentação concreta e operacional com elementos desse modelo (determinação do local pelo global, isotopias). O interesse do projeto depende, também, da medida em que a comunidade poderá se apropriar dele, apreciando um novo observatório de semântica interpretativa, como também, contribuindo, para ele, em termos de corpus ou propostas de exploração, infraestrutura ou documentação, e desenvolvendo novos conhecimentos compartilhados. Em termos do contexto e de seus interlocutores, o projeto Sittelle poderá ser, também, uma oportunidade para contribuir para a reflexão sobre novos modelos de edição digital,

13. Como fizemos para o corpus Zero experimental, que visa uma primeira visão geral rápida, optando por um formato pobre (txt), rápido de implementar, ao contrário da codificação TEI XML, que exige mais trabalho, embora permita, a médio prazo, capitalizar e compartilhar as melhorias realizadas .

visando aliar o acesso relevante aos escritos científicos e um modelo económico saudável, sustentável e equilibrado entre os editores.

Le projet est à la fois ambitieux et modeste : exigeant dans ses principes directeurs, enthousiaste dans ses perspectives scientifiques, mais posant d'entrée de jeu une démarche progressive, par étapes et ajustements, qui assume de communiquer de premières réalisations très partielles et expérimentales — un peu à la façon des méthodes dites « agiles » en développement informatique. D'ailleurs les Humanités numériquesne se prêtent pas à une division du travail entre tâches techniques et activité scientifique (BRADLEY, 2012) : le codage,méthodique et philologique, est une édition scientifique ; l'interrogation du corpus, construite et problématisée, est un parcours interprétatif.

O projeto é, ao mesmo tempo, ambicioso e modesto: exigente nos seus princípios orientadores, entusiasmado nas suas perspetivas científicas, mas que estabelec, desde logo, uma abordagem progressiva, por etapas e ajustamentos, que pressupõe comunicar conquistas iniciais muito parciais e experimentais – um pouco à semelhança de chamados métodos "ágeis" no desenvolvimento da Informática. Além disso, as Humanidades Digitais não se prestam a uma divisão de trabalho entre tarefas técnicas e atividade científica (BRADLEY, 2012): a codificação, metódica e filológica, é uma edição científica; a interrogação do corpus, construído e problematizado, é um percurso interpretativo.

O corpus ferramental, desenvolvido e compartilhado pelo projeto Sittelle, não substituirá as introduções pedagógicas ou os trabalhos de síntese sobre semântica interpretativa, muito menos o interesse e o prazer de ler os próprios textos: porque o corpus digital, naturalmente, entra em diálogo com a leitura atenta e contínua, apelando-se um para o outro mutuamente.

Este artigo deve muito ao apoio de François Rastier, sem o qual, obviamente, o projeto Sittelle não poderia ter nascido. Ele também se beneficiou do apoio paciente e especializado de vários colegas do laboratório IHRIM, em particular Alexei Lavrentiev, Serge Heiden e Matthieu Decorde da equipe TXM para discussões sobre codificação de texto, e Isabelle Treff e Nadine Pontal da divisão de Humanidades digital, para os primeiros referenciais fornecidos em questões jurídicas.

5. Referências bibliográficas

5.1. Bibliografia geral

BÉNEL, Aurélien. Archives numériques et construction du sens ou « Comment échapper au Web sémantique ? ».Gazette des archives, 245, 2017, p. 173-187.**https://doi.org/10.3406/gazar.2017.5524**

BRADLEY, John.No Job for Techies: Technical Contributions to Research in the Digital Humanities. In: Collaborative Research in the Digital Humanities,Farnham, Burlington: Ashgate, 2012, p. 11-.

BRUNET, Étienne. Tous comptes faits. Écrits choisis, tome 3. Questions linguistiques. Paris: Honoré Champion, 2016.

GUILLOT, Céline, HEIDEN,Serge, LAVRENTIEV,Alexei. Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. Diachroniques, 7, 2018, p.168-184.

GUILLOT, Céline, LAVRENTIEV, Alexei, RAINSFORD, Thomas, MARCHELLO-NIZIA, Christiane, HEIDEN, Serge. La « philologie numérique » : tentative de définition d'un nouvel objet éditorial. In:Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013), Section 13 : Philologie textuelle et éditoriale.Nancy: ATILF/SLR, 2017, p.143-154.

HÉBERT, Louis. Introduction à la sémantique des textes. Paris: Honoré Champion.

HEIDEN, Serge. Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex.In: Le poids des mots. Actes des 7es Journées internationales d'analyse statistique des données textuelles. Presses universitaires de Louvain, 2004, v.1, p. 577-588.

HEIDEN, Serge, MAGUÉ, Jean-Philippe, PINCEMIN, Bénédicte. TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In: Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010. Rome:EdizioniUniversitarie di LettereEconomiaDiritto, 2010, p. 1021-1031.

LEBART, Ludovic, PINCEMIN, Bénédicte, POUDAT, Céline. Analyse des données textuelles. Presses de l'université du Québec, 2019.

LEBART, Ludovic, SALEM, André. Statistique textuelle. Paris: Dunod, 1994.

MAYAFFRE, Damon. De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. Syntaxe & Sémantique, 9, 2008, p. 53-72.

MAYAFFRE, Damon. Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? In: Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification, Actes 2007 du colloque 2006. Toulouse: CALS-CPST (Colloque d'Albi Langages et signification-Centre pluridisciplinaire de sémiolinguistique textuelle), 2007a, p. 15-.

MAYAFFRE, Damon. Analyses logométriques et rhétoriques des discours. In: Introduction à la recherche en SIC. Presses universitaires de Grenoble, 2007b, p. 153-180.

NÉE, Émilie (dir.). Méthodes et outils informatiques pour l'analyse des discours. Presses universitaires de Rennes, 2017.

PINCEMIN, Bénédicte. Semánticainterpretativa y textometría. TópicosdelSeminario, 23, 2010, p. 15-55.

PINCEMIN, Bénédicte. Concordances et concordanciers. De l'art du bon KWAC. In: Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification, Actes 2007 du colloque 2006. Toulouse: CALS-CPST (Colloque d'Albi Langages et signification-Centre pluridisciplinaire de sémiolinguistique textuelle), 2007a, p. 33-42.

PINCEMIN Bénédicte. Introduction. Interprétation, contextes, codage. Corpus, 2007b, 6, p. 5-15.

PINCEMIN, Bénédicte. Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? In: Dépasser les sens iniques dans l'accès automatisé aux textes, Sémiotiques, 17, 1999, p. 71-120.

PINCEMIN, Bénédicte, HEIDEN, Serge, LAY, Marie-Hélène, LEBLANC, Jean-Marc, VIPREY, Jean-Marie. Fonctionnalités textométriques : Proposition de typologie selon un point de vue

utilisateur. In: Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010. Rome: EdizioniUniversitarie diLettereEconomiaDiritto, 2010, p. 341-353.

RASTIER, François. Data *vs*corpora. In: L'intelligence artificielle des textes — Des algorithmes à l'interprétation. Paris: Champion, 2021, p. 203-246.

RASTIER, François. Mesure et démesure. Quantité et qualité en linguistique de corpus. Le français moderne, 88 (1), 2020, p. 11-.

RASTIER, François. La mesure et le grain. Sémantique de corpus. Paris: Honoré Champion, 2011.

RASTIER, François. Arts et sciences du texte. Presses universitaires de France, 2001.

RASTIER, François. Sémantique interprétative. Presses universitaires de France, 1987.

RASTIER, François, CAVAZZA, Marc, ABEILLÉ, Anne. Sémantique pour l'analyse : de la linguistique à l'informatique. Paris : Masson, 1994.

RATINAUD, Pierre. Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In: JADT' 2018, Proceedings of the 14th international conference on statistical analysis of textual data. Rome, Italie: Universitalia, 2018,v.2, p. 616-6.

RATINAUD Pierre, DEJEAN Sébastien. IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre.In: Colloque Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009). Toulouse, 2009.

REINERT, Max. Alceste, une méthodologie d'analyse des données textuelles et une application: Aurelia, de Gérard De Nerval. Bulletin of SociologicalMethodology/Bulletin de Méthodologie Sociologique, 1990, 26(1), p. 24-54.

SCHMID, Helmut. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 1994.

VALETTE, Mathieu. Introduction — Pour une science des textes instrumentée. Syntaxe & Sémantique, 9, p.9-14.

5.2. Inventário de obras de pesquisa de François Rastier em língua francesa

I*déologie et théorie des signes*. La Haye: Mouton, 1971.

Essais de sémiotique discursive. Paris: Mame, 1973/1974.

Sémantique interprétative. Paris, Presses universitaires de France, 1987 (rééd. augmentées 1996, 2009).

Sens et textualité. Paris: Hachette, 1989 (rééd. augmentée Limoges: Lambert Lucas 2016).

Sémantique et recherches cognitives. Paris: Presses universitaires de France, 1991 (rééd. augmentée 2001).

Sémantique pour l'analyse — De la linguistique à l'informatique. En collaboration avec Marc Cavazza et Anne Abeillé. Paris: Masson, 1994.

Arts et sciences du texte. Paris: Presses universitaires de France, 2001.

Ulysse à Auschwitz— Primo Levi, le survivant. Paris: Éditions du Cerf, 2005.

La mesure et le grain — Sémantique de corpus. Paris: Champion, 2011.

Apprendre pour transmettre — L'éducation contre l'idéologie managériale. Paris: Presses universitaires de France, coll. Souffrance et théorie, 2013.

Saussure au futur. Paris: Les Belles-Lettres/Encre Marine, 2015.

Naufrage d'un prophète — Heidegger aujourd'hui. Paris: Presses universitaires de France, 2017.

Créer — Image, Langage, Virtuel. Paris-Madrid:Casimiro, 2016.

Heidegger, Messie antisémite — Ce que révèlent les Cahiers noirs. Lormont: Le bord de l'eau, 2018.

Faire sens — De la cognition à la culture. Paris: Classiques Garnier, 2018.

Mondes à l'envers — De Chamfort à Samuel Beckett. Paris: Classiques Garnier, 2018.

Exterminations et littérature — Les témoignages inconcevables. Paris: Presses universitaires de France, 2019.

5.3. Webgrafia

[Les URL ont toutes été vérifiées le 10 novembre 2021.]

Base de français médiéval. Céline GUILLOT-BARBANCE (dir.), École normale supérieure de Lyon, Université de Lyon, **http://bfm.ens-lyon.fr**.

Dictionnaire de sémiotique en ligne. Louis HÉBERT, Université du Québec à Rimouski, **http://www.semiotique.org**.

Guide d'application de la loi pour une République numérique – Art. 30, site Ouvrir la science. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, **https://www.ouvrirlascience.fr/guide-application-loi-republique-numerique-article-30-ecrits-scientifiques-version-courte/**.

Infrastructure éditoriale *OpenEdition*. OpenEdition Center, Unité de service et de recherche (USR 2004) du CNRS, d'Aix-Marseille Université, de l'EHESS et d'Avignon Université, **https://www.openedition.org**.

Text Encoding Initiative. **https://tei-c.org**.

Texto! Textes & Cultures. François RASTIER (dir.), Institut Ferdinand de Saussure, Programme Sémantique des textes, **http://www.revue-texto.net**.

TreeTagger — a part-of-speech tagger for many languages. Helmut SCHMID, Institute for Computational Linguistics of the University of Stuttgart& Center for Information and Language Processing of the University of Munich, **https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/**.

6. ANEXOS

6.1. Composition et réalisation du corpus Zéro

Le tableau suivant recense les 28 textes retenus pour le corpus Zéro (corpus test exploratoire) :

Code pour le corpus	***Titre du texte***	***Année de publication hors* Texto!**	***Adresse en ligne, après le préfixe* http://www.revue-texto.net/1996-2007/**
1996Concepts	La sémantique des textes : concepts et applications	1996	**Inedits/Rastier/Rastier_Concepts.html**
1996 PourSdT	Pour une sémantique des textes – Questions d'épistémologie	1996	**Inedits/Rastier/Rastier_PourSdT.html**
1996 Terme	Le terme : entre ontologie et linguistique	1995	**Inedits/Rastier/Rastier_Terme.html**
1996Themes	La sémantique des thèmes ou le voyage sentimental	1995	**Inedits/Rastier/Rastier_Themes.html**
1996 Transmission	Communication ou transmission ?	1995	**Inedits/Rastier/Rastier_Transmission.html**
1997 Defigements	Défigements sémantiques en contexte	1997	**Inedits/Rastier/Rastier_Defigements.html**
2001 Action	L'action et le sens – Pour une sémiotique des cultures	2001	**Inedits/Rastier/Rastier_Action.html**
2001 Bette	La Bette et la Bête – une aporie du réalisme	1992	**Inedits/Rastier/Rastier_Bette.html**
2001 Elements	Éléments de théorie des genres	-	**Inedits/Rastier/Rastier_Elements.html**
2001 Ling-de-style	Vers une linguistique des styles	2001	**Inedits/Rastier/Rastier_Ling-de-style.html**
2001 Semiotique	Sémiotique et sciences de la culture	2001	**Inedits/Rastier/Rastier_Semiotique.html**
2002 Acces	L'accès sémantique aux banques textuelles	2000	**Inedits/Rastier/Rastier_Acces.html**
2002 Genres	Genres et variations morphosyntaxiques	2001	**Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html**
2002 Macrosemantique	La macrosémantique	1994 (v1), 2002 (v2 en anglais)	**Inedits/Rastier/Rastier_Macrosemantique1.html**
2003 Semiotique-ontologie	De la signification au sens – Pour une sémiotique sans ontologie	1999 (en italien)	**Inedits/Rastier/Rastier_Semiotique-ontologie.html**
2004 Ontologies	Ontologies	2004	**Inedits/Rastier/Rastier_Ontologies.html**
2005 Classes-lexicales	De la sémantique cognitive à la sémantique diachronique : les valeurs et l'évolution des classes lexicales	2000	**Inedits/Rastier/Rastier_Classes-lexicales.html**
2005 Doxa	Doxa et sémantique en corpus – Pour une sémantique des « idéologies »	2005	**Inedits/Rastier/Rastier_Doxa.html**

Code pour le corpus	***Titre du texte***	***Année de publication hors* Texto!**	***Adresse en ligne, après le préfixe* http://www.revue-texto.net/1996-2007/**
2005 Enjeux	Enjeux épistémologiques de la linguistique de corpus	2005	**Inedits/Rastier/Rastier_Enjeux.html**
2005 Herm-et-ling	Herméneutique et linguistique : dépasser la méconnaissance	2003 (en allemand)	**Dialogues/Debat_Hermeneutique/Rastier_Herm-et-ling.html**
2005 Mesosemantique	Mésosémantique et syntaxe	1994 (v1), 2002 (v2 en anglais)	**Inedits/Rastier/Rastier_Mesosemantique.html**
2005 Microsemantique	La microsémantique	1994 (v1), 2002 (v2 en anglais)	**Inedits/Rastier/Rastier_Microsemantique.html**
2005 Saussure	Saussure au futur : écrits retrouvés et nouvelles réceptions. Introduction à une relecture de Saussure	-	**Saussure/Sur_Saussure/Rastier_Saussure.html**
2005 Semantique-cognitive	Sémiotique du cognitivisme et sémantique cognitive – Questions d'histoire et d'épistémologie	1993 + 1996	**Inedits/Rastier/Rastier_Semantique-cognitive.html**
2006 Borges	L'hypallage et Borges	2001	**Inedits/Rastier/Rastier_Borges.html**
2006 Formes-semantiques	Formes sémantiques et textualités	2006	**Inedits/Rastier/Rastier_Formes-semantiques.html**
2006 Inde	Saussure, la pensée indienne et la critique de l'ontologie	2002	**Saussure/Sur_Saussure/Rastier_Inde.html**
2006 Textes	Pour une sémantique des textes théoriques	2005	**Inedits/Rastier/Rastier_Textes.html**

Les textes ont été collectés sur le site Texto! en octobre 2021. Le contenu de la page HTML contenant le texte a été copié/collé dans un fichier .txt (encodé en UTF-8). Des transformations ont été appliquées en vue de l'intérêt des traitements textométriques.[14]

- Sont retirés : auteur(s) du texte, affiliation, mention de référence originelle, sommaire, bibliographie[15], annexes, indicateurs de pagination, abstract en anglais ; tout-à-fait en fin de page : le contact, la référence de l'article dans *Texto!*.
- Sont conservés : le résumé, les éléments de contenu textuels ou chiffrés issus de tableaux ou figures (mais pas les séquences de ponctuations ou

14. Essas transformações estão vinculadas ao formato de importação escolhido, do tipo texto bruto (.txt). Para o resto do projeto Sittelle, estamos considerando um formato de representação estruturada do tipo XML, que permitirá que certas informações sejam recodificados em vez de removidos, de acordo com princípios mais avançados de filologia digital. .

15. Pareceu-nos que a rede de autores citados permaneceu globalmente presente apesar da corrosão da bibliografia, graças às menções referentes à bibliografia ao longo do texto. Para o texto de Gêneros 2002 para o qual as referências bibliográficas utilizam um código, este código foi substituído pelo nome completo do autor.

de symboles), l'exergue (même en langue étrangère), les remerciements, les notes.

- Élimination des majuscules de mise en forme : lorsque le titre est en capitales, il est réécrit en typographie courante. Idem pour les intitulés de sections. Nous avons aussi réécrit les mentions des ruptures ou décrochements sémiotiques notées en majuscules par l'auteur (ex. ICI → 'ici'), mais à la réflexion ce n'était sans doute pas utile (on aurait pu garder ce système de notation original car il était régulier dans notre corpus).
- Variations typographiques de certains caractères : les tirets (cadratin, demi-cadratin et double tiret : --) sont tous convertis en tiret uniquesimple ; les guillemets doubles (et l'éventuel espace associé), présents en trois types différents (droits, chevrons français, américains) sont unifiés vers le guillemet double droit ; le guillemet simple et l'apostrophe, représentés de quatre façons différentes (droit, droit inversé, courbe dans les deux sens), sont tous réécrits en apostrophe droit ; les points de suspension représentés comme trois points successifs sont recodés avec le caractère dédié.
- On insère un espace entre le mot et l'appel de note qu'il porte, le cas échéant.
- Correction de quelques coquilles aperçues au passage (moins d'une dizaine).

6.2. Esquisse de la chaîne de traitement éditorial envisagée pour le corpus XML TEI

L'idée générale du projet Sittelleest de construire un corpus de textes au format**XML**selon les recommandations de la *TextEncoding Initiative* (**TEI**, **https://tei-c.org**), utilisant tous un même sous-ensemble simple des éléments de la TEI. L'explicitation de ce sous-ensemble, sous la forme d'un**schéma** dans un**document ODD**, vise une modélisation textuelle appropriée pour rendre compte de structures textuelles des écrits scientifiques de François Rastier pertinentes dans un contexte textométrique. La textométrie considère à la fois des informations de paramétragede certains calculs (ex. paragraphes pour définir des contextes, divisions du texte pour situer des résultats), de type de contenu pour des opérations de sélection (ex. relever du vocabulaire mais pas dans les bibliographies), mais aussi de présentation pour restituer des éléments de mise en page importants dans la lecture du texte (ex. soulignement).

Pour l'élaboration du schéma, nous comptons nous baser initialement sur notre connaissance préalable des textes de François Rastier (et des outils textométriques), et utiliser l'outil **Roma** (**https://roma2.tei-c.org**). Comme nous voulons une représentation légère, nous partirons d'un schéma minimal et préciserons des éléments nécessaires à ajouter. Nous nous attendons à ce que le schéma ne soit pas d'emblée satisfaisant mais se stabilise progressivement avec l'expérience de sa mise en œuvre, en effet ses premières utilisations sur notre corpus (codage puis interrogations textométriques) seront révélatrices de lacunes ou de choix à rectifier. Ce schéma devra d'ailleurs être documenté avec un **Guide d'encodage** (en principe intégré au document ODD), qui précisera en pratique l'interprétation des éléments TEI dans le contexte du corpus Sittelle, en se basant sur des cas concrets représentatifs. En effet, bien que très contrôlé formellement (liste limitée d'éléments disponibles, combinaisons possibles ou pas), le codage reste une activité fortement

interprétative et experte (c'est un travail d'édition scientifique). Nous pensons d'ailleurs que la mise au point du schéma et de sa documentation sera l'occasion d'une réflexion scientifique sur certains aspects de la textualité, de l'écriture scientifique et du corpus Rastier.

Pour les textes eux-mêmes, dans la mesure du possible, et avec le soutien de l'auteur, il s'agirait de partir de fichiers auteur en format **traitement de texte**. Si besoin, pour les fichiers plus anciens au format .doc, opérer une conversion vers .**docx** (au moyen du logiciel Microsoft Word),format qui en pratique assure une meilleure compatibilité avec l'outil de conversion vers XML TEI,**OxGarage** (**https://oxgarage.tei-c.org**), maintenu par la communauté TEI. La conversion automatique de docx à TEIP5 (profil « default ») produit un fichier XML TEI basé sur une interprétation standard des marques de mise en page, et pas nécessairement conforme au schéma XML défini pour Sittelle. L'édition du texte TEI avec un logiciel spécialisé pour XML, tel**Oxygen** (**https://www.oxygenxml.com** – outil commercial mais souvent disponible dans le monde de la recherche grâce à des licences partagées au niveau des universités ou des laboratoires), permettra d'ajuster le codage de façon assistée, pour le rendre conforme au schéma Sittelle et en adéquation avec les indications du Guide d'encodage. Certains pré-traitements généraux pourraient être rassemblés dans des **scenarios**Oxygen facilitant leur application méthodique.

Pour ce qui concerne l'**entête TEI**, on pourra envisager de gérer les informations qu'il contient dans un fichier unique de type**tableur** (cela permet une vue centralisée, plus facile à maintenir). En effet, on peut ensuite utiliser une feuille **XSL** pour alimenter puis mettre à jour automatiquement les entêtes dans le fichier TEI. La Base de français médiéval procède de cette façon (les métadonnées sont gérées dans une base de données).

L'ensemble des textes dans leurs différents états d'avancement dans la préparation du corpus, ainsi que les documents et ressources utilisés pour les traitements et plus généralement par le projet, pourront être gérés de façon partagée à l'échelle internationale via une solution de stockage de données sécurisée en ligne dédiée à la recherche en sciences humaines, le **ShareDocs** (**https://documentation.huma-num.fr/sharedocs-stockage/**) de l'infrastructure **Huma-Num** (**https://www.huma-num.fr**).