As palavras, o texto, os corpora e arquivo: o historiador face à linguística. Logometria e análise do discurso

Damon Mayaffre, Magali Guaresi, Laurent Vanni, Carlos Maciel

TEXTO DIGITAL
Revista de Literatura, Linguística e Artes

As palavras, o texto, os *corpora* e o arquivo: o historiador face à linguística. Logometria e análise do discurso.[1]

Les mots, le texte, les corpus et l'archive: l'historien face au linguistique. Logométrie et analyse du discours.

Damon Mayaffre[a]; Magali Guaresi[b]; Laurent Vanni[c]; Carlos Alberto Antunes Maciel[d]

a Universidade de Nice Sophia Antipolis, França – mayaffre@unice.fr
b Universidade de Nice Sophia Antipolis, França - magali.guaresi@gmail.com
c Universidade de Nice Sophia Antipolis, França - laurent.vanni@gmail.com
d Universidade de Nantes, França - carlos.maciel@unice.fr

Palavras-chave:
Logometria. Arquivo. História. Texto. Linguística. Corpus.

Resumo: Este artigo compreende duas grandes partes. São primeiramente reflexões sobre as inevitáveis relações entre a história e a linguística, na era da revolução digital, em que, ao lado do arquivo tradicional, encontramos as grandes bases de dados textuais e/ou arquivos históricos eletrônicos. Estas reflexões conduzem o leitor, pouco a pouco, para a questão fundamental do método, explicitada na segunda parte do artigo. Trata-se, aqui, de logometria, ou *medida do discurso*, e das diferentes ferramentas estatísticas e informáticas de que dispõem então o historiador e o linguista, mas também de leitura. Os exemplos são tirados da história política francesa.

Keywords:
Logométrie. Archive. Histoire. Texte. Linguistique. Corpus.

Abstract: Cet article comprend deux grandes parties. Nous avons d'abord des réflexions sur les rapports qui se nouent inévitablement entre l'histoire et la linguistique, à l'ère de la révolution numérique dans laquelle, aux côtés de l'archive traditionnelle, nous trouvons désormais les grandes bases de données textuelles et/ou les archives historiques électroniques. Ces réflexions conduisent peu à peu le lecteur vers la question fondamentale de la méthode, qui fait l'objet de la deuxième partie de l'article. Il s'agit alors de logométrie, ou *mesure du discours*, et des différents outils statistiques et informatiques dont dispose l'historien et le linguiste, ainsi que de lecture. Les exemples donnés sont issus de l'histoire politique française.

[1] Traduzido por Carlos Alberto Antunes Maciel – BCL (UMR 7320 / NuPILL**.** Título do original francês: ***Les mots, le texte, les corpus et l'archive : l'historien face au linguistique. Logométrie et analyse du discours***. Os autores são membros do Laboratório BCL (*Bases Corpus et Langage* / Bases, Corpus e Linguagem), UMR (*Unité Mixte de Recherche* / Unidade Mixta de Pesquisa) 7320, CNRS (*Centre National de Recherche Scientifique* / Centro Nacional de Pesquisa Científica) – Universidade de Nice, e membros do grupo logometria. **Damon MAYAFFRE** foi diretor do Laboratório BCL, de que o grupo logometria é parte integrante, e é membro do grupo Politext – de estudo de textos políticos.

Há mais de vinte anos que foi publicado o principal artigo de Antoine Prost « Les mots », in *Pour une histoire politique* (Prost 1988), e pretendemos, com esta contribuição, prolongar a reflexão sobre a posição que o material linguístico *(les mots* – as palavras, *la phrase* – a frase, *le texte* – o texto) ocupa nas ciências históricas, no que se refere à análise do *discurso* político na história e nas relações entre a história e a linguística, que inevitavelmente se constroem quando desta espécie de embate corpo a corpo que o historiador trava com os *corpora* ou com o *arquivo*.

Para dar resposta, como num eco, ao nosso grande predecessor, dois aspectos serão aqui abordados: a situação epistemológica na França que, por um lado, no início do século XXI, aparece de forma bem diferente daquilo que ela era no final do século precedente e, por outro lado, a evolução do método proposto por Prost, que chamaremos agora de *logometria* (isto é, *medida do discurso*), o qual, graças à generalização dos recursos da era digital e ao desenvolvimento das ferramentas estatísticas e informáticas, foi progressivamente melhorando e até mesmo banalizou-se.

Apesar de gratuita, a antiga objeção feita por Roland Barthes, citada por Régine Robin na sua obra *Histoire et Linguistique*, pioneira nesta matéria, será levada em consideração : “Il est constant qu’un travail qui proclame sans cesse sa volonté de méthode soit finalement stérile. / *Muitas vezes acontece que um trabalho em que muito se proclama a necessidade do método acabe finalmente por ser estéril*.” (Robin 1973 : 8).

Não vamos então somente aqui ilustrar a proposta metodológica com os resultados concretos obtidos pelos historiadores no campo da história política contemporânea, assim como seria possível também fazer no campo da história moderna, medieval ou antiga[2]. O objetivo é o de mostrar a fertilidade de uma (hiper)leitura dos textos, com ferramentas, e não somente com um tipo de leitura que se poderia chamar de selvagem, não somente intuitiva, mas controlada, com a condição de bem determinar o objetivo que ela se dá (não simplesmente pretender objetivar *o* sentido dos textos, mas objetivar percursos de leitura ou percursos interpretativos

[2] É em história moderna e sobre o discurso revolucionário que a lexicometria, com os modernistas de Saint-Cloud (Régine Robin, Jacques Guilhaumou, Annie Geffroy, etc.), produziu bons resultados. Para a história medieval, vamos ler, por exemplo, as obras de Jean-Philippe Genet e, mais recentemente: Aude Ma*yret, Une Angleterre entre rêve et réalité. Littérature et société dans l’Angleterre du XIXe siècle*, Paris, PuS, 2007. No que se refere à história antiga, o *LASLA – Laboratoire d'Analyse Statistique des Langues Anciennes* (Universidade de Liège) foi, deste ponto de vista, pioneiro. Ver, na bibliografia, as referências às obras de Dominique Longrée e Sylvie Mellet.

verificáveis e reprodutíveis) de que se deve bem determinar a condição (considerar o que se refere à linguagem não como um vetor transparente da história ou como um simples *medium* para o historiador, mas como um elemento constitutivo da história e como um objeto de pesquisa *à part entière)*.

ANDANDO EM VOLTA DO *TEXTO;*

ENTRE OS CO*RPORA E* O ARQUIVO.

1 RENEGOCIAÇÃO DISCIPLINAR, REVIRAVOLTA HERMENÊUTICA E REVOLUÇÃO DIGITAL

Longe dos grandes sistemas explicativos como o marxismo ou o estruturalismo, que constituem o pano de fundo da aproximação que se produz entre a história e a linguística, nos anos 1960-1980, descrita por Antoine Prost, a paisagem científica atual está marcada por alguns elementos que se nos impõem no que se refere às relações dos pesquisadores em ciências humanas – e particularmente o historiador – com o que é textual ou linguístico. Trataremos dos dois mais importantes, aqueles que são inevitáveis dentro do que é cotidianamente praticado pelos pesquisadores do século XXI: a revolução digital e a reviravolta hermenêutica nas Ciências Políticas e Sociais.

1.1 A REVOLUÇÃO DIGITAL

Através, voluntariamente, de um belíssimo atalho, o antropólogo britânico Jack Goody resume a história da humanidade, na qual a invenção da escrita foi determinante, a bem pouca coisa:

> Le passage […] du geste au langage est à la base de notre humanité, et le passage d'une culture du manuscrit à une culture de l'imprimerie est à la base de la modernité. / *A passagem […] do gesto para a linguagem está na base da nossa humanidade, e a passagem de uma cultura do manuscrito para uma cultura da imprensa está na base da nossa modernidade* (Goody 2007 : 163)

Goody mostra-nos, assim, como, depois da emergência da linguagem, e depois também do nascimento decisivo da escrita, que fez com que o homem entrasse na história (*versus* a pré-história), a imprensa, em meados do século XV, inventou a modernidade, participou daquilo a que se deu o nome de Humanismo e do Renascimento, contribuiu para que se fizesse a

Reforma – a religião de um Livro divulgado –, e para que surgissem também as Luzes – a filosofia das enciclopédias universais. Mais recentemente, o desenvolvimento das ciências, assim como o da imprensa, particularmente no século XIX, deve ser visto como diretamente dependente da demultiplicação e da circulação da escrita, tanto a escrita erudita quanto a escrita vulgar, que fez com que Gutenberg surgisse e, com ele, os seus sucessores.

É sob o prisma deste dinamismo histórico multisecular que cabe que se tente medir a revolução digital atual, cujas repercussões na civilização são e serão provavelmente ainda mais importantes do que a própria invenção da imprensa em meados do século XV: o nascimento da hipermodernidade; do hipertexto digital a partir do texto impresso; da hiperleitura e do hiperleitor depois da leitura e do leitor.

Com a expressão revolução digital queremos, antes de mais nada, designar a mudança histórica que representa um sustentáculo para a cultura humana, isto é, a passagem, de forma acelerada e generalizada, do papel para o digital. Num certo prazo, e muito rapidamente, se considerarmos a escala da história das sociedades, é o conjunto do texto (e da imagem e do som) que será produzido, formatado, armazenado e transportado através do digital[3]. No final do ano de 2011, depois de somente 6 anos de trabalho, *Google Books* (books.google.fr, para os Franceses) anunciou que tinha já processado e indexado, em versão integral, mais de 6% de todos os livros que foram até hoje publicados no planeta, o que representa cerca de 600 bilhões de palavras (dos quais 45 bilhões cabiam então à língua francesa – hoje já há cerca de 100 bilhões em língua francesa com relação a um total que já passa de um trilhão, em diferentes línguas)[4]. Na França, a BNF (*Bibliothèque Nationale de France)* pôs à disposição dos usuários, em 2012, no seu sítio *Gallica*, 1,8 milhão de documentos, entre os quais

[3] Não se trata para nós de fazer uma qualquer profecia no que se refere ao desaparecimento do papel; Mc Luhan já foi bastante imprudente quando fez isso nos anos 1960. De modo transitório, muito pelo contrário, o digital favoreceu o papel. Por exemplo, a troca de arquivos e as tiragens individuais conduziram a uma multiplicação das publicações personalizadas, feitas diretamente na impressora de cada um. O único campo onde a relação entre o papel e o digital se fez em benefício deste, em termos de ruptura e de exclusividade, é o da produção: não há hoje nenhum doutorando e nenhum escritor escrevendo ainda a sua obra diretamente no papel; a leitura hoje é uma leitura mixta papel/digital, mas a escrita erudita – como também a vulgar – SMS, correio eletrônico – tornou-se quase que exclusivamente digital.

[4] Cf. Jean-Baptiste Michel *e alii*, "Quantitative Analysis of Culture Using Millions of Digitized Books", *SciencExpress*, 16 de dezembro de 2010 (http://www.librarian.net/wp-content/uploads/science-googlelabs.pdf), ou as informações fornecidas em linha por *Google Books* na rubrica « about Google books ». Sabemos que *Google Books* assinou, já no ano de 2006, uma convenção com 5 grandes bibliotecas (Nova Iorque, Universidades de Harvard, do Michigan e de Stanford e a bodleiana de Oxford) para o resgate sistemático dos acervos; a empresa fez depois disso novos acordos, com outras bibliotecas, para vir a ser, em princípio, universal.

encontramos milhares de manuscritos do acervo histórico da Biblioteca. A imprensa diária dos países desenvolvidos e as revistas científicas estão desde já também em linha, em portais eletrônicos ricamente elaborados (*Revue.org* ou *Persee.fr* na França para as revistas em Ciências Humanas e Sociais). Vastos bancos ou armazéns são criados para a difusão de teses, de esboços ou ainda de documentos de trabalho: na França, *HAL-archives-ouvertes* tinha já 300.000 documentos científicos que podiam ser consultados já a partir de abril de 2014, dos quais 45.000 em Ciências Sociais e Humanas, etc.

Assim, se é verdade que este novo estado de coisas é, considerada a sua dimensão, impressionante, é também verdade que o que há de mais extraordinário nisso tudo está ainda num outro aspecto da questão. Estes bilhões de documentos podem, com efeito, ser consultados imediatamente por todo e qualquer internauta, dentro do seu próprio centro de pesquisa, dentro da sua própria casa ou ainda a partir de um ponto qualquer do planeta, através dos satélites, com um telefone celular ou com um tablet digital. Com cerca de 3 bilhões de seres humanos (em grande parte no mundo ocidental), e muitos mais de lá para cá (2,79 bilhões estavam por exemplo conectados às redes sociais em dezembro de 2016), a revolução de que aqui se trata tem tremendos efeitos e fica fora das ambições desta contribuição.

No que se refere aos estudos históricos, as consequências são sem precedente. Durante muito tempo condenada pela raridade das fontes, seja porque elas são mesmo pouco numerosas (tal como acontece com o cartulário antigo) seja porque estavam dificilmente disponíveis (como podia acontecer com aquela série de arquivos que era preciso exumar num município longínquo qualquer), o historiador tem hoje condições de reunir, no mesmo instante e com somente alguns cliques, uma quantidade imensa de documentos de trabalho. Na França, a maioria dos arquivos departamentais e municipais está, com efeito, sendo digitalizada. Os pequenos arquivos departamentais de Lozère ou de Mayenne, por exemplo, oferecem desde já ao historiador internauta, sem nenhuma restrição, nenhum filtro e nenhuma forma de pagamento, a totalidade dos registros ditos *BCS* (Batismos, Casamentos, Sepulturas), desde o decreto de Villers-Cotterêts[*], e a totalidade do registro civil pós-revolução. Em outros lugares,

[*] Nota do Tradutor: Através do Decreto de Villers-Cotterêts, de 1539, o Rei François 1er (Francisco I) fez do francês a língua oficial dos procedimentos administrativos e da corte na França, desalojando assim o latim, que era usado até então (artigos 110 e 111). Este decreto é o mais antigo texto legislativo ainda em vigor na França.

são as minutas dos diferentes cartórios da época moderna que podem ser consultadas livremente. Em Troyes, são todos os mapas do século XII da Abadia de Clairvaux[**] que estão em linha. Na Suíça, entre inúmeros outros exemplos, temos o caso extraordinário do acervo da Abadia de Saint Gall, do século VIII, que está sendo também agora digitalizado (www.sg.ch). E assim vem acontecendo com as fontes (textuais) de que deve poder dispor habitualmente o historiador.

Para o historiador contemporâneo e mais ainda para o historiador do tempo presente, a situação é mais radical ainda e chega a atingir proporções vertiginosas, já que a realidade social e política do mundo se deixa agora apreender por um fluxo ininterrupto de documentos digitalizados, a que o utente tem acesso sem mesmo precisar sair de casa: são discursos, Diários Oficiais, documentação oficial, índices econômicos, levantamentos demográficos, imprensa, fóruns, arquivos de correio eletrônico, blogues ou ainda redes sociais[5]. Foi no período de 2010-2011 que cerca de 90.000 documentos diplomáticos secretos ou relatórios de campo do exército dos Estados Unidos no Afeganistão (alguns até mesmo pertencentes à área da defesa) foram disponibilizados por um sítio (*WikiLeaks*) e logo divulgados pelas agências de imprensa. Concretamente, trabalhamos com os discursos presidenciais franceses: eles podem ser baixados no dia mesmo em que foram proferidos, em formato digital – e unicamente neste formato – do sítio oficial do Palácio do Eliseu; sem falar do blogue pessoal do ex-presidente Nicolas Sarkozy, da página Facebook de François Hollande (ou de Obama)[***], e de outros mais, através dos quais os presidentes se dirigem diretamente aos seus concidadãos, por via eletrônica, no século XXI. Da mesma maneira, os sítios credenciados pelo governo – www.ladocumentationfrancaise.fr et www.vie-publique.fr – divulgam milhares de documentos (discursos sindicais, discursos ministeriais, relatórios, decretos, portarias, textos de leis, etc.) cujo armazenamento em papel será feito talvez algum dia, mas este dia é por enquanto incerto e imprevisível. Entre outras riquezas que ali estão para a história política, a base "discursos" da Vida-Pública dá acesso a mais de 150.000 textos de

[**] Nota do Tradutor: A Abadia de Clairvaux ou Abadia de Claraval (*Clara Vallis* em latim, foi um mosteiro cisterciense fundado em 1115 por Bernardo de Claraval (futuro São Bernardo) e está localizada em Ville-sous-la-Ferté, no departamento de Aube, na França. Bernardo de Claraval vinha da Abadia de Cister e ali chegou com alguns monges (*Citeaux*).

[5] Será que podemos imaginar uma história das mentalidades do ano 2000 sem que sejam analisadas as redes sociais digitais (Twitter, Facebook, etc.) ? Será que o historiador vai poder, por exemplo, um dia, rastrear a história das revoluções da Tunísia e do Egito, no início de 2011, sem analisar a importância das trocas de informações feitas através da Internet, que ultrapassaram o sistema oficial de informações dos governos de Ben Ali e de Moubarak ?

[***] Nota do Tradutor: Também os discursos dos presidentes brasileiros estão disponíveis.

líderes políticos, de responsáveis associativos ou de simples personalidades destes trinta últimos anos.

O historiador (do contemporâneo) encontra-se, assim, face a um continente desconhecido: administrar a abundância, e até mesmo, podemos dizer, o infinito – a quantidade quase excessiva de textos instantaneamente disponíveis – no lugar da raridade anterior – e a pequena quantidade de textos efetivamente consultáveis; administrar a virtualidade digital no lugar do papel bem material; administrar a dinâmica do Web quando o arquivo tradicional era mais estável ou pelo menos mais estático; administrar também o caráter imediato e talvez até mesmo efêmero (dos fluxos) e, ao mesmo tempo, o caráter patrimonial (dos acervos)[6].

Uma coisa é certa: o arquivo textual digital de hoje – o Web – , na sua dimensão quase ilimitada, evolutiva, interativa, proteiforme, não pode ser visto ingenuamente como um sábio e fiel vetor para o historiador, capaz de o conduzir simplesmente na direção de uma realidade histórica objetiva; ele aparece, na sua dinâmica, na sua ubiquidade, no seu emaranhado infinito, como parte constituinte desta história que também graças a ele se escreve. O arquivo Web, imenso, movediço, nunca idêntico a ele mesmo, onipresente na sua atualidade cidadã, deixa de ser descritivo ou referencial e torna-se, de modo evidente, performativo com relação à história que conta ou prescritivo no que se refere à história que constrói. Nunca antes as afirmações de Jacques Guilhaumou e Denise Maldidier, do final do século passado, pareceram tão pertinentes para descrever o arquivo textual digital contemporâneo :

> C'est que l'archive n'est pas le reflet passif d'une réalité institutionnelle : elle est, dans sa matérialité et sa diversité mêmes, mise en ordre par son horizon social. L'archive n'est pas un simple document où se puisent les référents ; elle s'offre à une lecture qui découvre des dispositifs, des configurations signifiantes. / *É que o arquivo não é o reflexo passivo de uma realidade institucional: ele está, na sua materialidade, como na sua própria diversidade, ordenado pelo seu horizonte social. O arquivo não é um simples documento onde vamos buscar referentes; ele presta-se a uma leitura que descobre dispositivos, configurações significantes.* (Guilhaumou *et alii* 1994 : 92)

Em suma, o arquivo Web do historiador (do contemporâneo) é, antes de mais nada, um

[6] Pode-se compreender que as práticas do historiador acabem por encontrar um importante desafio, na montante, da parte do arquivista ou do bibliotecário. Como é que vão ser armazenados, no longo prazo, os documentos gerados em formato digital (isto é, aqueles que só existem em formato digital)? Alguns precedentes desagradáveis mostram desde já a fragilidade do arquivo eletrônico. O que fazer, por exemplo, com as mensagens de correio eletrônico que constituem hoje o essencial das trocas internas de mensagens da administração e da diplomacia estadunidenses?

discurso (ou um horizonte discursivo): o discurso de uma época sobre ela mesma, que não figura a história mas a configura, que não diz a história mas a co-fabrica. E, como discurso, merece ser tratado não intuitivamente, mas, antes, com os fundamentos metodológicos que a análise do discurso vem propondo nestes últimos trinta anos, durante os quais os pesquisadores refletiram sobre o teor linguístico das realidades históricas. Particularmente, vamos sugerir ao leitor a obra de Jacques Guilhaumou, *Discours et événement,* que encerra, em 2006, uma reflexão de várias décadas sobre o lugar que devem ocupar as realidades linguísticas na sua relação com a história, sobre a performatividade histórica do discurso, sobre o funcionamento do "evento discursivo". A obra milita, com Koselleck, por uma história linguística dos conceitos que afastam o historiador da sua pretensão de mexer diretamente com os referentes (as " coisas" ou os "fatos" históricos) através das palavras. Sobretudo, ele evidencia uma mediação necessária do discursivo na produção do evento histórico: os acontecimentos produzem-se para os que dele são contemporâneos, mas é ao historiador que cabe fazer deles a leitura e os compreender, numa dimensão que não é para eles atributiva, mas, antes, constitutiva. Vamos, então, repetir mais uma vez: o arquivo, o discurso, a linguagem não contam a história; antes, dela participam. Eles não são a expressão ingênua e a narração passiva de uma história que se faria sem eles: são a condição dela – condição cognitiva particularmente.

Quanto ao *corpus de trabalho*[7] que resulta do arquivo, está claro que ele aparece menos do que nunca como um dado, mas como um objeto construído, fruto de uma escolha determinada pela razão e feita dentro do arquivo Web, isto é, naquele oceano de documentos que podem ser instantaneamente solicitados. Aqui, esta afirmação, neste século XXI digital, passa a ter um eco especial:

> On puise dans ce que Jean Dubois appelait « l'universel du discours », c'est-à-dire dans la "totalité des énoncés d'une époque, d'un locuteur, d'un groupe social". Découpage arbitraire, à partir d'intérêts, de thèmes, de savoir… / *Recorremos àquilo que Jean Dubois chamava de "universal do discurso", isto é, o "conjunto dos enunciados de uma época, de um locutor, de um grupo social". Divisão arbitrária, a partir de interesses, de temas, de saber...* (Guilhaumou *et alii* 1994 : 76)

Ora o "universal do discurso" foi durante muito tempo um fantasma que ninguém esperava

[7] O *corpus*, pela sua transversalidade, impôs-se nestes últimos anos como o objeto central das Ciências Humanas e Sociais: não há estudo sem que seja constituído um *corpus* de trabalho. Desde 2001, uma revista está consagrada a este objeto de estudo: *Corpus* (http://corpus.revues.org/)

nem mesmo entrever e, no próprio seio deste inacessível horizonte, o *corpus* não era o resultado de uma "divisão arbitrária" em função de uma problemática, mas, antes, uma aceitação do pior resultante de uma seleção drástica imposta pela modicidade das fontes existentes ou materialmente consultáveis. Hoje, sem mesmo imaginar ingenuamente que o Web veicula "o conjunto dos enunciados de uma época, de um locutor, de um grupo social", pretendemos que, pela primeira vez na história das ciências históricas, ele aproxima-se um pouco disso; a nossa pretensão é a de que, pelo menos, ele dá uma aproximação muito superior à dos arquivos existentes, de papel[8]. Concretamente, isto sim, no sítio do Palácio do Eliseu estão disponíveis, e instantaneamente acessíveis, em formato texto e em vídeo, todas as intervenções públicas do Presidente (discursos, mensagens, votos desejados, telegramas de felicitações, conferências de imprensa, alocuções, conversas informais através do blogue). Nada falta, por exemplo, da palavra pública de Nicolas Sarkozy ou de François Hollande, ou muito pouco. Tudo está à disposição (quando os discursos do Presidente Pompidou, na década de 1960, nem mesmo foram antes disponibilizados e não estão ainda, em uns tantos casos, disponíveis). No *corpus* de Jacques Chirac, porque já era então o caso, como no *corpus* de Nicolas Sarkozy ou no de François Hollande, construídos, a base é a exaustividade ou o "universal" da palavra presidencial que pode ser solicitada através de um simples copiar e colar[9].

Cabe aqui, então, pôr em relevo, já finalizando, o mais importante. A passagem do papel para o digital não é somente uma mudança técnica ou material de suporte da cultura humana que tem como consequência uma demultiplicação e uma circulação do escrito bem mais poderosas do que aquelas que a invenção da imprensa provocou; demultiplicação e circulação que têm como consequência, hoje, em história, a modificação da definição de *arquivo textual* com base nas nossas práticas científicas. A passagem para o digital provoca, sobretudo, uma evolução ainda mais importante, quase antropológica, da nossa relação íntima e essencial com

[8] Lembremos aqui quais são os limites inerentes à história. Trabalhamos com os traços textuais (o arquivo) que nos foram legados pelo passado. Estes traços só cobrem uma parte irrisória da realidade dos que dela são contemporâneos. Como o arquivo de papel, o arquivo digital também não pode, nem mais nem menos, ter a pretensão de poder descrever esta realidade, mas o Web permite multiplicar (por 10? 100? 1 000?) a existência e as possibilidades de consulta destes traços.

[9] Não cabe que entremos aqui nos detalhes técnicos. Melhor do que copiar e colar, as ferramentas de aspiração e de gestão dos fluxos da Internet dão ao pesquisador a possibilidade de recuperar automaticamente os novos documentos que, cotidianamente, são disponibilizados pela Secretaria do Palácio do Eliseu. Um sistema de alerta avisa o analista quando houve uma modificação no arquivo Web e, em consequência, no seu *corpus* de trabalho. O arquivo torna-se assim exaustivo e dinâmico, e pode também assim ser consultado, com a dinâmica que o caracteriza.

o texto e com a leitura (isto é, com as condições do conhecimento erudito, e também com as condições daquela relação que, informada pela linguagem, nós temos com a realidade histórica): este é o verdadeiro sentido da revolução digital de que hoje se trata.

O que é um texto no momento em que o digital se impõe? O que significa ler um documento digital? O que é a textualidade no ecrã ou tela dos nossos computadores? Como é que se lê (isto é, como é que se compreende) o Web, o arquivo, o *corpus* hoje? Será que o escrito digital produz "efeitos de sentido" e, esperemos, uma "mais-valia de sentido" que o leitor desconhecia até então?

Revolução: arrancando-o do *scriptorium* e da biblioteca, "retirando-o" do lugar onde estava fixado, no livro, no papel, ou no maço que até aqui representavam o seu único suporte, o digital definitivamente *desnaturalizou* e *desmaterializou* o texto (o que fez com que se voltasse em parte a uma prática que existiu da Antiguidade até a Idade Média, quando os textos, copiados manualmente, comentados e interpretados, não eram fac-símiles mecânicos). Antes havia uma tendência para naturalizar o texto, assimilando-o, através da sua fixidade material, ao seu suporte físico tradicional (a página, o livro), mas a filologia digital tornou hoje evidente a sua *artefatualidade*[10]. Pluralidade de formatos e de códigos, escolhas múltiplas e individuais de visualização, multiplicação dos níveis de etiquetagem, de anotações ou de enriquecimento, circulação, sem limites, até à ubiquidade, através de um arquivo anexado ou partilhado, etc.: tudo se combina para melhor sublinhar a partir de então a volatilidade, a virtualidade ou a relatividade do texto, a sua dimensão artefatual, convencional ou cultural (isto é, não natural). Como o *corpus*, o texto não aparece mais como um dado; ele torna-se uma construção, e o historiador, de mãos dadas com o linguista, muito vai ganhar se tentar desconstruí-lo, isto é, se tentar analisá-lo sem inocência e com método, e tal como ele é: um objeto complexo que só faz sentido na sua complexidade, e não um material elementar, transparente ou intuitivo.

[10] Antecipando no que se refere à reflexão *infra*, cabe citar Dominique Legallois: « Bien sûr, il faut mentionner l'hypertextualité liée à la mutation numérique du texte, qui oblige à une reconsidération de l'unité textuelle et du texte lui-même : en tant qu'ensemble de possibilités de parcours, le texte devient alors *une unité virtuelle* » (*Langages*, n° 163, 2006, p. 7). O conjunto do desenvolvimento que segue deve muito aos pensadores do texto e da f*ilologia digital*: François Rastier, *Arts et Sciences du texte*, Paris, PUF, 2001, chp. III, "Philologie numérique', p. 73-97 ; Jean-Marie Viprey, "Philologie numérique et herméneutique intégrative", dans Jean-Michel Adam et Ute Heidmann (éds.), *Sciences du texte et analyse de discours*, Genebra, Slatkine, p. 51-68 ; e a nossa tese: Damon Mayaffre, *Vers une herméneutique matérielle numérique. Corpus textuels, logométrie et langage politique* (de habilitação para dirigir trabalhos de pesquisa, ou HDR, defendida na Universidade de Nice no dia 30 de abril de 2010 ; http://tel.archives-ouvertes.fr/tel-00655380).

Tecnicamente sobretudo – e remetemos, desde já, aos métodos de tratamento apresentados a seguir – , vamos claramente indicar que a informática *deslineariza* o texto (como começou a fazer, no primeiro século da nossa era, o *codex* (o códice) com relação ao rolo. Durante muito tempo considerado unicamente como uma *sequência* linguística contínua, com um início, um meio e um fim, o texto digital apresenta-se ao contrário como uma *rede* linguística descontínua, com as suas conexões, aparelhamentos e remissões. O texto podia ser visto, antes de mais nada, através da sua *progressão* ou do seu *desenvolvimento*, do começo até o ponto final, mas ele se apresenta a nós agora também com o seu entrelaçamento e as suas remissões internas, com os seus ecos semânticos ou de coocorrências, com as suas conexões hipertextuais que apontam também para o lado de fora: os nossos textos transformaram-se em hipertextos nos quais as unidades (as palavras, por exemplo) estão materialmente (através de notas ou de elementos aos quais se fixam, numa espécie de ancoragem, de chaves de índices, de relações hipertextuais) conectadas entre elas e para além delas[11] para produzir efeitos de sentido e percursos de leitura até então desconhecidos.

Porque é deste modo que o próprio ato de leitura acabou por ser modificado: trata-se aqui do ponto central desta contribuição, que vai propor ao historiador todo um arcabouço metodológico para dar uma moldura a esta nova leitura (uma hiperleitura) deste novo objeto (o hipertexto ou texto digital).

Durante muito tempo *linear*, a leitura passou a ser também *tabular* e *reticular* segundo os teóricos da matéria: à vontade de ler o *corpus* na sua continuidade acrescentou-se hoje uma vontade de extrair as *tábuas* ou os *quadros* de frequência, por exemplo (as palavras mais utilizadas ou os índices de frequências), de solicitar *listas* (o vocabulário organizado na forma de um dicionário ou um índice alfabético), de pôr em relevo redes lexicais privilegiadas (as coocorrências, os temas privilegiados, os campos lexicais), de atualizar as relações intratextuais (as remissões hipertextuais *via* uma palavra entre os textos dentro do *corpus*) e as relações intertextuais (as remissões hipertextuais fora do *corpus*). Penetraremos, assim, no *corpus* menos talvez através da primeira palavra, da primeira frase, para uma leitura corrente[12], do que através de uma palavra-chave, de uma entrada do índice ou das

[11] Cf. *infra* e Christian Vandendorpe, *Du papyrus à l'hypertexte. Essai sur les mutations du texte et de la lecture*, Paris, La Découverte, 1999.

[12] Como também há muito tempo renunciamos à ideia de dar muita importância à primeira letra da

concordâncias (com todas as frases que contêm uma palavra-polo). Mais longe ainda, é a proporção ou distribuição quantitativa de tal ou tal forma, de tal ou tal categoria gramatical, de tal ou tal tempo verbal, de tal ou tal sequência sintática que poderá vir a ser um modo de leitura. O historiador estadunidense do livro e da leitura, Robert Darnton, declarou que estava maravilhado com as mudanças em curso e, em termos simples, concluiu:

> [Les lecteurs modernes liront] de façon horizontale, verticale ou diagonale, selon les directions ouvertes par les liens électroniques. / *[Os leitores modernos vão ler] de modo horizontal, vertical ou diagonal, segundo as direções abertas pelas conexões eletrônicas.* (Darnton 2011 : 180)

Temos, quanto a nós, chamado a atenção para a existência, em vários livros e artigos (ver bibliografia), da complementaridade entre a leitura digital e a leitura ocular tradicional: a leitura hipertextual como complementar com relação à leitura textual clássica[13], a leitura quantitativa complementar com relação à leitura qualitativa, a leitura paradigmática do computador complementar com relação à leitura sintagmática humana, a leitura rizomática e segmentária complementar com relação à leitura contínua ou sequencial, e, tal como foi acima indicado, a leitura tabular e reticular complementar com relação à leitura linear.

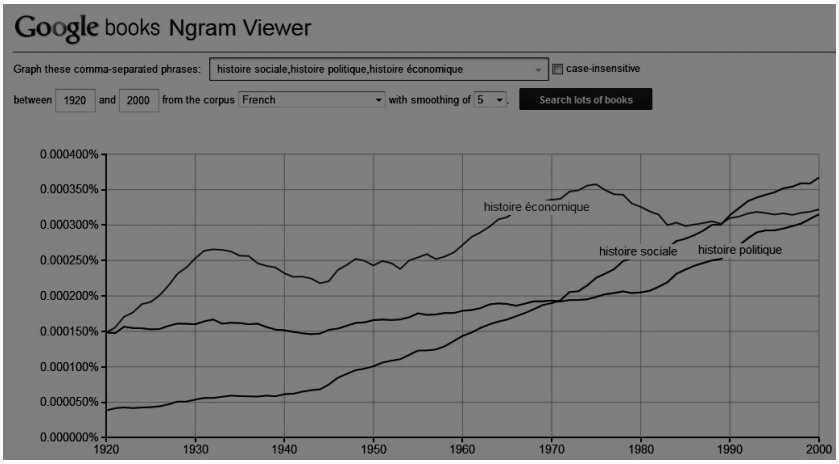
Resumindo, para um suporte digital, uma leitura digital. Os documentos tão numerosos de hoje não têm vocação para serem lidos somente pelo olho humano – isto, diante da imensidão, torna-se em alguns tantos casos impossível. Eles prestam-se, no entanto, a processamentos (semi)automáticos. Os motores de pesquisa – antes de falar das ferramentas logométricas mais aperfeiçoadas – são os primeiros que devem ser mencionados; ninguém hoje deixa de fazer uso deles. Com os seus algoritmos muito potentes, eles tornam-se indispensáveis como ferramentas de leitura de um arquivo Web incomensurável. Eles permitem ao mesmo tempo extrair informações (de uma forma ainda grosseira) mas, sobretudo, entrar num *corpus* que é, no entanto, gigantesco, através de uma qualquer palavra que terá servido para começar a busca. Para além destes simples motores de pesquisa – e agora já falando um pouco mais das ferramentas logométricas – observemos enfim, por exemplo, que há 1 trilhão de palavras

primeira palavra, já desprovida de iluminuras.

[13] E vamos tratar de explicitar imediatamente este ponto essencial : o primeiro movimento dos programas informáticos de logometria consiste em indexar sistematicamente todas as unidades do *corpus* (mais ou menos como se faz com os nomes próprios numa obra impressa). Assim, todas as unidades (todas as palavras, por exemplo) podem tecnicamente ser localizadas com um simples clique, no próprio *corpus*, ou solicitadas (com as suas frases), quantificadas, classificadas, etc.

extraídas de alguns milhões de livros que se encontram agora disponíveis no *Google Books*, o qual, por outro lado tem, desde 2010, uma ferramenta de tratamento lexical que permite observar a distribuição quantitativa de uma palavra ou de uma sequência de palavras num período de vários séculos (ilustração 1).

Figura 1. Frequências de utilização dos sintagmas "história social", "história política", "história econômica" no *Google Books* (100 bilhões de palavras em francês).



1.2 A REVIRAVOLTA HERMENÊUTICA DAS CIÊNCIAS HUMANAS E SOCIAIS

A reviravolta hermenêutica de que queremos agora falar tem relações com a revolução digital, mas é anterior a ela e é *a priori* independente.

Por reviravolta hermenêutica, não vamos aqui evocar as obras de Schleiermacher e de Dilthey, nem as de Heidegger ou de Gadamer, mas só, livremente, dois elementos que atualmente se combinam na pesquisa em Ciências Humanas e Sociais.

Trata-se, primeiramente, da tomada de consciência global de que não somente a interpretação é a nossa condição de estar no mundo, como também de que ela é modal nas práticas cotidianas em Ciências Humanas e Sociais: a ideia e o gesto interpretativos são parte integrante, em todas as partes e sempre, do nosso percurso científico. No que se refere às disciplinas históricas e ao debate que aqui se faz a propósito do material linguístico que o historiador é obrigado a apreender, cabe que se diga claramente – nos termos daquilo que é o primeiro princípio da hermenêutica – que a interpretação dos textos ou do arquivo apresenta-se como essencial *para a própria determinação e descrição que se quer deles fazer;* e vamos então afirmar que o arquivo (e particularmente o arquivo Web), considerada a sua riqueza e a

sua diversidade, não pode ser abordado, visto e descrito sem ser, ao mesmo tempo, interpretado ou compreendido. Afirmemos, ainda, que os recursos interpretativos do nosso arquivo universal a interpretar encontram-se, nos termos de uma dificuldade hermenêutica bem identificada, nele mesmo; e vamos, a partir daí, repetir aquela constatação problemática admitida pela filosofia: a linguagem (aqui, aos poucos e por aproximação, o arquivo Web, o *corpus* que dele resulta ou o texto) tem a característica de ser, ao mesmo tempo, a condição e o meio necessários para a sua própria interpretação.

Afirmar, então, com Szondi, por exemplo[14], que *descrição* e *interpretação* do objeto não são pontos que possam ser separados um do outro (o que quer dizer, em consequência, que as nossas práticas não são nem objetivantes nem formalizantes, mas, antes, fundamentalmente hermenêuticas) conduz-nos a reconsiderar a fronteira disciplinar entre a história e a linguística em análise do discurso.

Já não temos mais, de modo mais ou menos esquizofrênico, dois textos: o texto que o linguista se encarregaria de descrever e o texto que o historiador se encarregaria de interpretar (no melhor dos casos, com base na descrição feita pelo outro). Há, a partir de agora, um só objeto, e um projeto interdisciplinar global e unificador que tem por objetivo descrever e interpretar e, ao mesmo tempo, receber e compreender o texto num mesmo movimento, *captar* (materialmente) e *captar* (intelectualmente), em conjunto. Esta reflexão metodológica envolve de forma muito abrangente a questão das realizações linguísticas em história, tal como ela foi acima mencionada: textos, discursos, linguagem não são ferramentas, veículos ou fontes frias e transparentes para o historiador, mas, antes e enquanto tais, objetos de análise. E cabe, então, precisar aqui que recebendo, tal como ela é, a dimensão das realizações linguísticas da história humana, renunciamos não somente a abordar o texto ou a palavra como simples "transmissores" evidentes como renunciamos também, de forma mais sutil, às abordagens co-variacionistas cujo objetivo era simplesmente o de constatar o paralelismo ou o isomorfismo entre, por um lado, a prática discursiva, e, por outro lado, a realidade histórica objetiva que a observa: entre o discurso e a realidade histórica não há nenhuma co-variação, há co-substancialidade; o *texto* (linguístico) não está sendo iluminado "de fora" pelo *contexto* (histórico): o texto é parte constitutiva do contexto, assim como as realizações linguísticas são

[14] Peter Szondi, *Introduction à l'herméneutique littéraire. De Chladenius à Schleirmacher*, Paris, Cerf, 1989.

parte constitutiva da história.

Ao lado destas reflexões gerais, a reviravolta hermenêutica designa uma reflexão mais precisa sobre o texto e o *corpus* textual como objetos, conduzida pelos linguistas por volta do ano 2000; mas esta reflexão parece ser mesmo interessante, em primeiro lugar, para o historiador. Dominadas, durante muito tempo, pelos linguistas da Língua que, dentro da tradição chomskiana, trabalhavam de modo introspectivo sobre o sistema linguístico, as ciências da linguagem requalificaram recentemente a Palavra e revalorizaram os dados linguísticos (discursos efetivamente pronunciados, palavras efetivamente ditas, textos realmente produzidos)[15]. Além disso, a generalização dos grandes *corpora* – e encontramos novamente aqui a revolução digital – favoreceu o desenvolvimento de uma linguística mais empírica baseada em dados (os dados atestados são sempre mais confiáveis, mais numerosos, mais representativos e até mesmo exaustivos) e não mais com base no *exemplum* (as frases que o gramático apresentava como exemplos e que eram muitas vezes inventadas)[16]. Mais adiante ainda, autores como Jean-Michel Adam ou François Rastier puderam finalmente afirmar que a única produção linguística realmente produzida, em situação, pelo locutor, não é nem a palavra sozinha, nem a frase isolada, mas, antes de mais nada, o discurso na sua forma empírica observável, o texto. Os textos, e os *corpora* textuais, tornaram-se, então, no início dos anos 2000, um objeto não somente legítimo mas inevitável para a linguística.

Ora – e é aqui que encontramos a reviravolta interpretativa –, quando o texto tinha a sorte – que raramente teve – de ser levado em consideração pelo linguista, era para imaginar uma gramática de textos ou uma semântica formal complexantes e nem sempre produtivas para o historiador. Mas hoje este ponto de vista está radicalmente transformado. A linguística textual atual reivindica, com François Rastier, uma longa tradição "retórico-hermenêutica" (*versus* uma tradição "lógico-gramatical") (Rastier 2011 : 17) e afirma que o texto e o seu sentido são, antes de mais nada, objetos de interpretação. Com a afirmação segundo a qual só há sentido com a interpretação (e não com a formalização), a semântica textual se dá como objetivo estabelecer percursos interpretativos credíveis e reprodutíveis dentro dos próprios *corpora* textuais, apoiando-se, sem dúvida e rigorosamente, nas unidades linguísticas do texto, mas

[15] Esta revalorização muito deve à extraordinária descoberta de manuscritos inéditos de Saussure, É*crits de linguistique générale*, Paris, Gallimard, 2002. Ler (Monte et Philippe (dir.) 2014).

[16] Bernard Laks, "Pour une phonologie de corpus", *Journal of French Language Studies*, n° 18, 2008, p. 3-32.

concedendo à contextualização e à exigente liberdade interpretativa o primeiro lugar. Com um centro de gravidade disciplinar deslocado, é este o programa hermenêutico que pode ser aproveitado pelos historiadores.

É em todos os casos exatamente nestes termos que vamos definir o objetivo do tratamento logométrico dos textos em história: dar objetividade aos percursos de leitura; dar objetividade aos percursos interpretativos; nunca pretender provar ou formalizar o sentido – porque os textos são sempre polissêmicos e o sentido, plural, é sempre o fruto de uma interpretação – mas objetivar ou enquadrar caminhos interpretativos reprodutíveis, falsificáveis, balizados por paradas explícitas do processamento.

Antoine Prost tinha claramente sentido estas coisas ao falar do valor probatório e, mais ainda, ao falar do valor heurístico da lexicometria. Hoje, é preciso insistir com relação a este valor heurístico ou hermenêutico. O computador, com uma leitura diferente, faz perguntas diferentes para alimentar/construir de outra maneira a interpretação. Com as suas capacidades de armazenamento, de triagem, de classificação, de navegação, o computador não prova nada, mas permite ver, ou melhor, interpretar elementos do *corpus* que ficaram invisíveis: ele baliza e põe ferramentas nos caminhos da interpretação.

2 DO TRATAMENTO LOGOMÉTRICO DOS DISCURSOS

É assim que, enquadrado pela revolução digital e, epistemologicamente falando, pela reviravolta hermenêutica, o método logométrico pode, com grandes pinceladas, ser descrito e ilustrado.

2.1 DEFINIÇÃO E FORÇA

Devemos entender por *logometria*, uma *lexicometria* de segunda geração que prolonga o método nascido na década de 1980 e que estende todas as suas práticas exploratórias e estatísticas até então limitadas ao léxico (*lexi*) a todas as unidades do discurso (*logos*): as palavras gráficas, mas também os lemas, as categorias morfossintáticas, os traços gramaticais, as cadeias sintáticas, etc., serão assim levados em consideração. Lexicometria, logometria ou ainda textometria foram já várias vezes apresentados nos últimos anos e, além das contribuições de Robin ou de Prost, já citadas, é também nas contribuições mencionadas na

bibliografia, de Brunet, Lebart e Salem, Lemercier e Zalc, Genet, Guilhaumou, Labbé, Marchand, Mayaffre, Tournier, etc., que vamos encontrar outros subsídios.

A força do método vem do processamento, que combina uma leitura que é, na sua essência, qualitativa, mas com o auxílio da informática (pesquisa documental, extração de passagens e de trechos, volta controlada e sistemática ao texto através de pontos de apoio ou de ancoragens, navegação hipertextual no interior de grandes *corpora*) e uma leitura que é, na sua essência, quantitativa (análise estatística do vocabulário ou dos traços lexicais, método de classificação dos textos, atualização dos elementos recorrentes ou significativos, localização das temáticas para o cálculo das coocorrências, formalização de redes de palavras que tenham algum tipo de parentesco no texto, etc.): trata-se, então, de *ler* mas também de *contar*. Para além desta complementaridade qualitativo-quantitativo, a logometria articula um tratamento micro ou local do texto (análises das unidades no seu contexto imediato – o sintagma, a frase, o parágrafo – , localização exata destas unidades nos grandes *corpora*, e volta aos textos para que se possa fazer deles uma leitura tradicional) e um tratamento macro ou global (visão sintética do *corpus*, descrição quantitativa geral, localização sistemática das distribuições e regularidades linguísticas, descrição da arquitetura geral dos textos).

O que pretendemos fazer aqui é ilustrar as principais ferramentas, mostrando os percursos de leitura que iniciam e remetendo à bibliografia para todos os detalhes. As ferramentas apresentadas têm todas elas a vantagem de terem sido integradas aos melhores programas informáticos de logometria, tais como Hyperbase, Iramuteq, Lexico ou TXM.

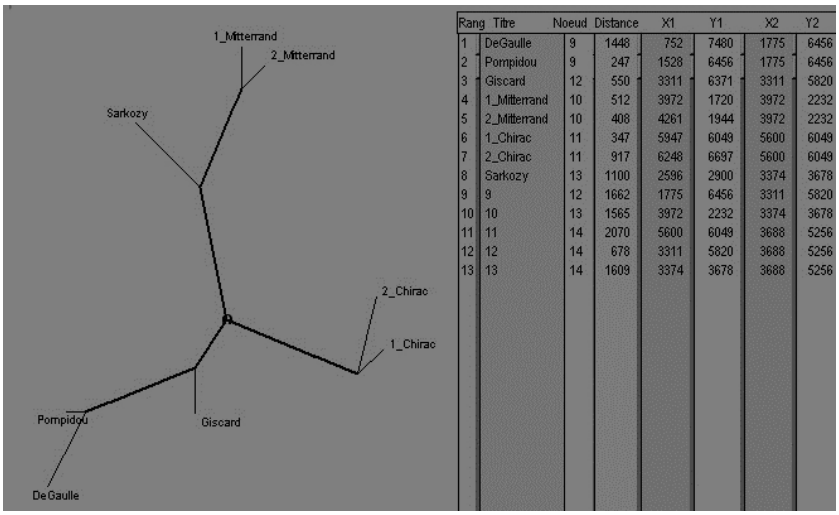
2.2 CLASSIFICAÇÃO (E REPRESENTAÇÃO EM ÁRVORES)

Vamos tratar aqui do grande *corpus* dos discursos presidenciais franceses de 1958 a 2012, que reúne os principais discursos dos presidentes, de Charles de Gaulle a Nicolas Sarkozy (700 discursos reunidos em 8 “textos” para um número igual de mandatos presidenciais). O primeiro movimento de busca é, então, quase sempre de tipo emergencial ou indutivo, e as primeiras questões gerais serão: quais são, globalmente, os textos (isto é, os presidentes) que mais se aproximam ou mais se distanciam um do outro? A que genealogia discursiva pertencem, por exemplo, os discursos de Nicolas Sarkozy ?

É para responder a este tipo de perguntas gerais que os estatísticos propuseram cálculos de

distância intertextual ou de conexão lexical, com representações em árvores: cf. historicamente Muller (1977), e, depois, no número 2 da revista *Corpus* (2002), e em Brunet (2011), ou ainda no nosso ensaio de vulgarização (Mayaffre e Luong – 2003) (ilustração 2).

Figura 2. Distância intertextual calculada a partir dos lemas. Representação em árvore segundo o método Luong (com Hyperbase 10.0 – 2017).



Rang	Titre	Noeud	Distance	X1	Y1	X2	Y2
1	DeGaulle	9	1448	752	7480	1775	6456
2	Pompidou	9	247	1528	6456	1775	6456
3	Giscard	12	550	3311	6371	3311	5820
4	1_Mitterrand	10	512	3972	1720	3972	2232
5	2_Mitterrand	10	408	4261	1944	3972	2232
6	1_Chirac	11	347	5947	6049	5600	6049
7	2_Chirac	11	917	6248	6697	5600	6049
8	Sarkozy	13	1100	2596	2900	3374	3678
9	9	12	1662	1775	6456	3311	5820
10	10	13	1565	3972	2232	3374	3678
11	11	14	2070	5600	6049	3688	5256
12	12	14	678	3311	5820	3688	5256
13	13	14	1609	3374	3678	3688	5256

Na árvore, podemos ver que os discursos de Charles de Gaulle e de Georges Pompidou estão próximos, e que esta proximidade é um sinal de que temos aí um falar gaulliano prototípico. No outro extremo da árvore, temos os discursos dos dois mandatos de sete anos de François Mitterrand, que estão reunidos: apesar da evolução do presidente socialista durante os seus dois mandatos, um falar mitterrandiano existe. Este falar encontra-se no lado completamente oposto com relação ao falar gaulliano.

É no âmbito de um tal quadro que é interessante constatar que os discursos de Nicolas Sarkozy estão mais próximos dos discursos de François Mitterrand do que dos discursos de Charles de Gaulle, como também de Jacques Chirac.

Como vai na contramão da intuição, esta constatação, que vamos voluntariamente deixar sem uma interpretação finalizada, é heurística. Quais são as causas de uma tal aproximação (que é neste caso lexical) entre Sarkozy e Mitterrand? Como interpretar esta proximidade inesperada? Neste caso, outras ferramentas (cf. *infra*) serão necessárias para levar mais adiante o percurso interpretativo que começou por baixo (*buttom up*), isto é, sem nenhum *a priori*, sem nenhuma hipótese e sem nenhum pré-julgamento político.

Com resultantes probantes no nosso exemplo com 8 textos (ou 8 mandatos presidenciais), as ferramentas de classificação trabalham com potência total quando os *corpora* têm muitos textos e seu objetivo é a datação de textos desconhecidos ou a atribuição de autoria. Assim, num *corpus* com 100 discursos, Calvet e Véronis (2008) puderam classificar e identificar os discursos de Nicolas Sarkozy escritos por Henri Guaino e os que não tinham sido escritos por ele: é a influência "secreta" do conselheiro em comunicação que foi então posta em relevo. Labbé e Monière (2003) utilizam a mesma ferramenta para decifrar o discurso governamental canadense, francês e do Québec, durante cerca de meio século.

2.3 CARACTERIZAÇÃO

A logometria – anteriormente chamada de lexicometria – é sobretudo conhecida pelas suas capacidades de caracterização dos textos. A ferramenta mais utilizada é o cálculo das especificidades, que intervém depois de um tratamento exaustivo para a localização das formas características de uma parte do *corpus* com relação ao conjunto (Lebart e Salem 1994).

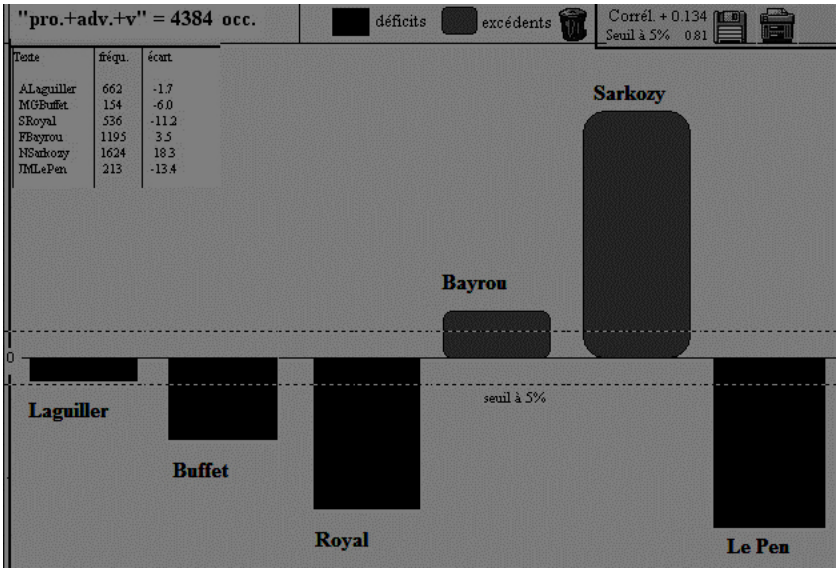
Assim, no mesmo *corpus* em que estávamos anteriormente, vamos poder saber quais são os lemas característicos de Charles de Gaulle (com relação aos outros presidentes) e exprimir numericamente (neste caso, com um desvio reduzido) o grau da sua hiperutilização pelo fundador da Quinta República[****]: *peuple*/povo (+22), *algérien*/argelino (adj.) (+19), *destin*/destino (+16), *État*/Estado (+16), *univers*/universo (+15), *régime*/regime (+15), *atomique*/atômico (+14), *nation/*nação (+13), etc. (Mayaffre 2004).

Graças a este cálculo, cuja vantagem é, antes de mais nada, a de ser sistemático, todas as palavras foram examinadas, sem exclusão nenhuma e sem nenhum interesse específico determinado *a priori*. Podemos, a partir daí, pôr em relevo um termo ou uma unidade linguística de que se pretende representar a distribuição no *corpus*. No *corpus* da campanha presidencial de 2007, o tratamento das cadeias linguísticas permite perceber que uma das mais fortes características do candidato Nicolas Sarkozy é a hiperutilização da estrutura "pronome

[****] Nota do Tradutor: A Quinta República começa em 1958, quando é adotada, como o nome indica, a quinta constituição republicana do País. A Quinta República teve, até aqui, os seguintes Presidentes: Charles de Gaulle, Georges Pompidou, Valéry Giscard d'Estaing, François Mitterrand, Jacques Chirac, Nicolas Sarkozy e Emmanuel Macron (que está atualmente em exercício).

+ advérbio + verbo", que, em francês, é a marca da negação. A sua distribuição no *corpus* é representada assim (ilustração 3):

Figura 3. Distribuição da estrutura "pronome + advérbio + verbo" no *corpus* da campanha eleitoral de 2007 (de Hyperbase, versão 10.0 – 2017).



É através desta recorrência estatística que toda a ideologia reacionária de Nicolas Sarkozy pode ser melhor estudada (Mayaffre 2007 e 2012). Explicitamente, em 2007, Sarkozy decidiu "liquidar a herança de maio de 1968"[17]. Este programa passa pela afirmação de uma autoridade que controla e voluntariamente contraria os slogans do tipo "gozar juntos" ou "é proibido proibir" de maio de 1968 (e a sua proposta de "trabalhar mais" aparece, assim, como uma outra faceta desta reação ao espírito de 1968[*****]). E esta autoridade do pai severo da nação exprime-se repetindo no seu discurso frases com a estrutura negativa "pronome + advérbio + verbo" : "*je ne veux pas*/eu não quero", "*je n'accepterai pas*/eu não aceitarei", "*il ne faut pas*/não se deve", "*vous ne pouvez pas*/vocês não podem", "*vous n'avez pas le droit*/vocês não têm o direito", etc.

[17]Nicolas Sarkozy, discurso de Montpellier, 3 de maio de 2007.

[*****] Nota do Tradutor: Candidatos, em 2007, na ordem em que aparecem no gráfico de distribuição: Arlette Laguiller (*Lutte Ouvrière* / Luta Operária), Marie-George Buffet (Partido Comunista), Ségolène Royal (Partido Socialista), François Bayrou (MoDem – *Mouvement Démocrate* / Movimento Democrata), Nicolas Sarkozy (UMP – *Union pour un Mouvement Populaire* / União para um Movimento Popular), Jean-Marie Le Pen (*Front National* / Frente Nacional).

Faz-se aqui alusão à palavra de ordem que mais parece ter marcado a campanha presidencial de Nicolas Sarkozy, em 2007: "*travailler plus pour gagner plus* / trabalhar mais para ganhar mais", e às grandes greves e manifestações, de estudantes e de trabalhadores, que, em maio de 1968, muito marcaram a França e obrigaram o governo a fazer uma reviravolta política.

De um modo geral, cabe lembrar aqui uma das principais regras do tratamento logométrico: o tratamento faz-se no âmbito de uma semântica diferencial (o sentido cria-se pela diferença) e de uma estatística endógena. O *corpus* constitui uma norma semântica e estatística, e é pelo contraste com relação a esta norma (aqui, no caso, o *corpus* de todos os candidatos às eleições presidenciais de 2007), que uma parte do *corpus* (no caso, Nicolas Sarkozy) se caracteriza. Foi assim que Pascal Marchand (2007) pôde caracterizar o vocabulário de cada Primeiro Ministro francês desde 1958 com relação ao discurso ministerial médio do período.

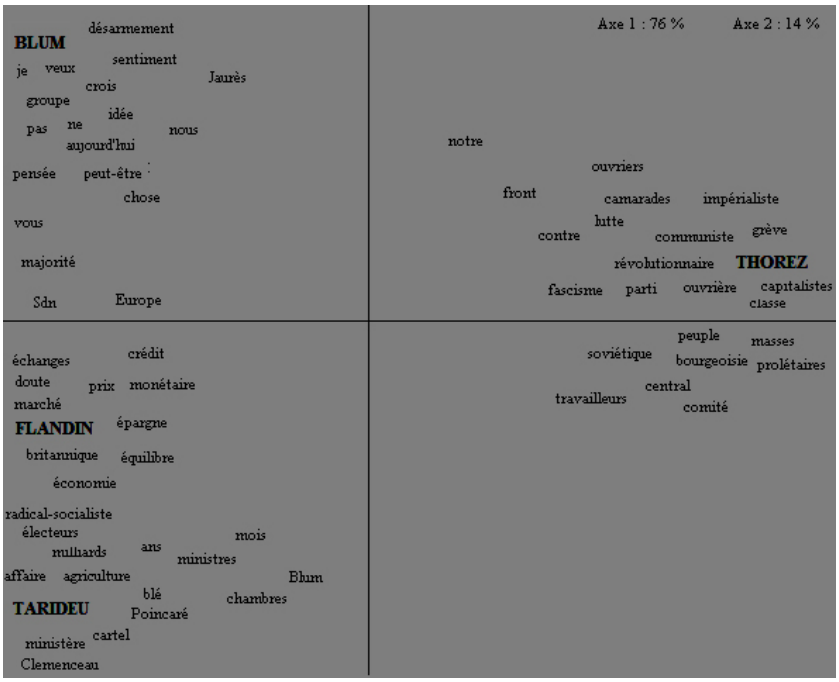
2.4 CORRESPONDÊNCIAS (ANÁLISE FATORIAL DE CORRESPONDÊNCIAS)

É no ponto de intersecção da ideia de processamento global classificatório com a caracterização dos textos que encontramos a principal ferramenta logométrica, que continua a ser a análise fatorial de correspondências (AFC) (Benzécri 1973 ; Prost 1974).

Vejamos, então, o caso de um *corpus* do período que se situa entre as duas grandes guerras, que é composto de quatro líderes políticos (Maurice Thorez, Léon Blum, Pierre-Etienne Flandin e André Tardieu[******]), e onde encontramos 60 palavras principais e muito frequentes. É possível "fazer corresponder" os líderes e as palavras, isto é, calcular as afinidades lexicais de uns e de outros, de modo a obter a seguinte representação (ilustração 4):

[******] Nota do Tradutor: Maurice Thorez (Partido Comunista), Léon Blum (Partido Socialista, membro da SFIO, *Section Française de l'Internationale Ouvrière* – Secção Francesa da Internacional Operária), Pierre-Etienne Flandin e André Tardieu (que pertenceram ambos à então chamada Aliança Democrática) são os quatro grandes líderes políticos do período considerado.

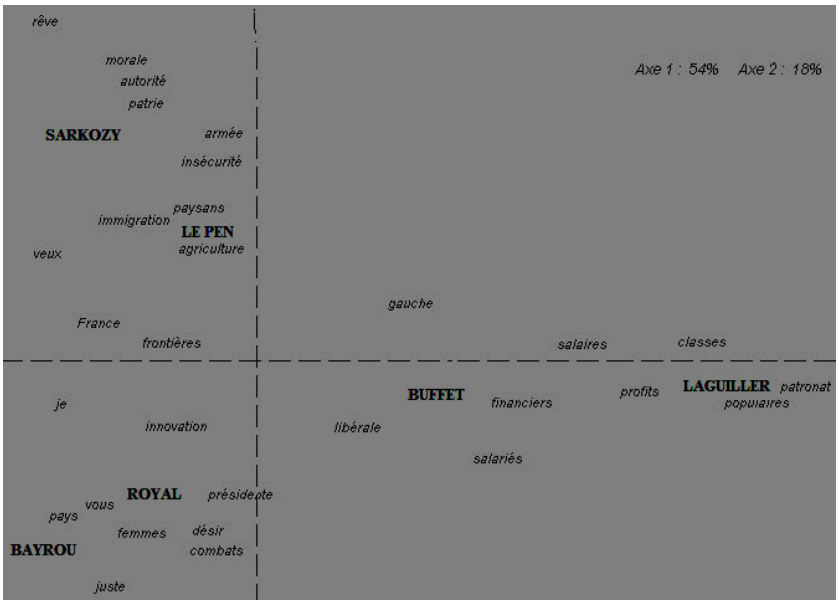
Figura 4. AFC de 60 palavras num *corpus* do período que se situa entre as duas grandes guerras (de Hyperbase, versão 10.0 – 2017).



A partir daí, é possível concluir que o discurso comunista ficou marginalizado (sozinho do lado direito do gráfico – com Maurice Thorez, grande líder comunista de então) e que, inversamente, o discurso socialista (Léon Blum), já entre as duas grandes guerras, se tinha aproximado do discurso republicano ou burguês; já naquela época então Léon Blum, o grande líder socialista (do lado esquerdo do gráfico), tinha deixado de lado não somente "*les prolétaires*/os proletários" ou "*le prolétariat*/o proletariado" mas também "*les ouvriers*/os operários", "*le peuple*/o povo" ou "*les travailleurs*/os trabalhadores" (Mayaffre 2000).

Da mesma maneira, a nuvem de pontos da campanha de 2007 é sugestiva e faz com que Nicolas Sarkozy apareça perto de Le Pen (no mesmo quadrante – ilustração 5).

Figura 5. AFC de 30 palavras durante a campanha eleitoral de 2007 (de Hyperbase, versão 10.0 – 2017).

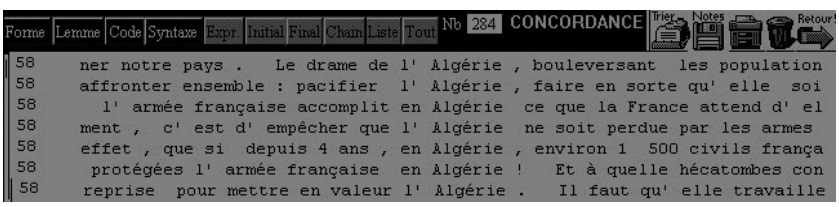


2.5 CONTEXTUALIZAÇÃO

As ferramentas de co(n)textualização são essenciais em logometria. São o objeto necessário para que as ferramentas quantitativas possam funcionar. Graças a elas, é possível ler o texto de modo tradicional, mesmo se considerarmos que elas acrescentam a esta leitura a sistemacidade e a exaustividade. Para além da hipertextualidade generalizada (todas as palavras do *corpus*, no programa, estão vinculadas a um índice, podem ser consultadas e representam um número igual de ancoragens que permitem voltar ao texto), a ferramenta mais conhecida é o *concordancier* (Pincemin 2006).

Graças a esta ferramenta, podemos solicitar, sob a forma de uma lista (com triagem, como é cabível), todas as passagens que contêm uma palavra ou uma unidade, para uma leitura controlada, e, depois, a partir desta lista, voltar, com um simples clique, para um contexto mais abrangente (ilustração 6):

Figura 6. Concordância de *Algérie*/Argélia nos discursos de Charles de Gaulle, em 1958 (de Hyperbase, versão 10.0 – 2017).

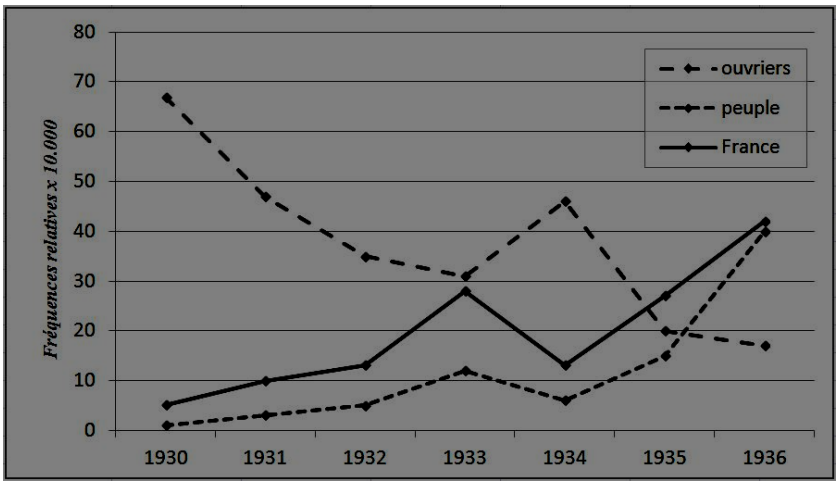


Por trás desta ferramenta, cuja utilização está hoje banalizada, está a posição da citação, discutida por Prost (1988), que mudou. Como a leitura passou a ser exaustiva (com todas as passagens), os trechos que vierem a ser utilizados para a demonstração passam a ser probantes, e não mais somente ilustrativos; o seu caráter significativo poderá agora ser julgado quando antes somente era possível duvidar da sua representatividade e do modo de seleção aleatório da sua seleção pelo tratamento manual.

2.6 DESCRIÇÃO CRONOLÓGICA

Para o historiador, uma menção particular deve ser feita com relação às ferramentas de descrição cronológica. A AFC é frequentemente empregada, assim como coeficientes de correlação cronológica capazes de localizar palavras cujo uso aumenta/diminui com o passar do tempo num *corpus* diacrônico, e remetemos, neste caso, às obras de André Salem ou de Dominique Labbé. Mas uma simples frequência relativa já diz, muitas vezes, bastante coisa (ilustração 7).

Figura 7. Frequências de utilização de *France* (França)
peuple (povo) e *ouvriers* (operários) nos discursos de Maurice Thorez (1930-1936).



Como as palavras representadas são significativas da "grande reviravolta" do Partido Comunista Francês (PCF) entre as duas grandes guerras, temos de forçosamente constatar que a mudança começa precocemente, desde 1931, 1932 e 1933. Mas esta precocidade pode não ter sido claramente posta em relevo por causa da anomalia de 1934, que fez com que se voltasse para a tendência de fundo. Em suma, damos uma interpretação e uma cronologia diferentes do nascimento do *Front populaire*/da Frente Popular, que convém assim ir buscar

mais longe, na mudança de vocabulário (dos *ouvriers*/operários ou dos *prolétaires*/proletários para o *peuple*/povo), no início da década de 1930 (Mayaffre 2000).

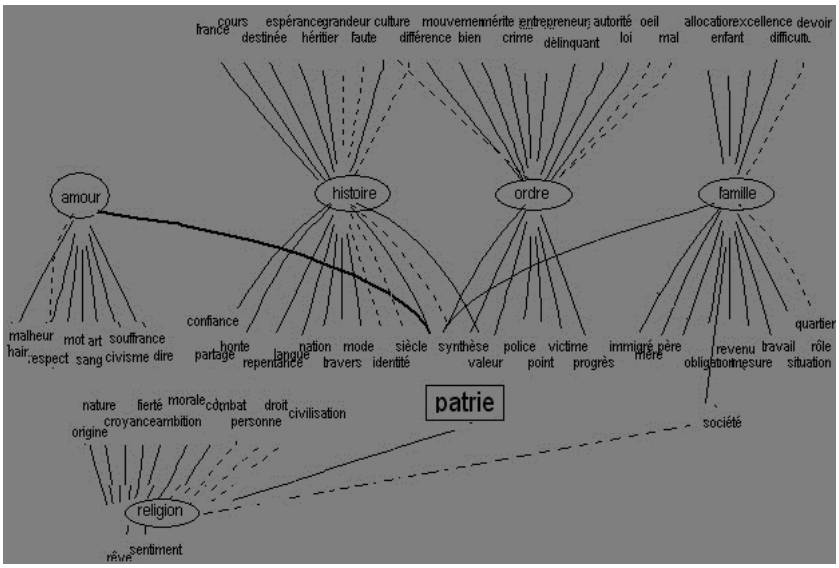
2.7 TEMATIZAÇÃO E COOCORRÊNCIAS

É preciso, então, parar um pouco para melhor analisar as coocorrências, que representam, ao mesmo tempo, o tratamento mais rico, o mais abundante e o mais interessante para o historiador; é também o tratamento que mais progrediu nos últimos anos (Mayaffre 2014). E, como este canteiro de obras ainda está aberto, ele vai então ser utilizado aqui através de uma única ilustração que não tem naturalmente a pretensão de esgotar a pluralidade dos enfoques possíveis nesta matéria.

Se o tratamento das ocorrências chega, por vezes, até às fronteiras em que se encontra com o semântico (a palavra sozinha não quer dizer nada), o tratamento das coocorrências (a palavra nas suas relações com as suas vizinhas estatísticas) permite chegar aos temas abordados e ao sentido do discurso (tal colmo acontece com toda uma parte da pesquisa em que se pode entrar, por exemplo, por Heiden e Lafon (1998) ou Mayaffre (2008, 2014)).

Vejamos o caso da palavra *patrie*/pátria nos discursos de Nicolas Sarkozy durante a campanha presidencial de 2007. A sua frequência de utilização faz logo com que cada um possa sentir a sua importância, mas o que é que isto exatamente significa? O tratamento das suas coocorrências (depois, por iteração, o estudo dos co-ocorrentes dos co-ocorrentes) permite ver que há neste caso 5 acepções diferentes e, um pouco mais adiante, ver também em parte a ideologia do candidato (ilustração 8).

Figura 8. Co-ocorrentes de *patrie* (pátria) nos discursos de Nicolas Sarkozy (2007) (com Hyperbase versão 9.0 – 2014).



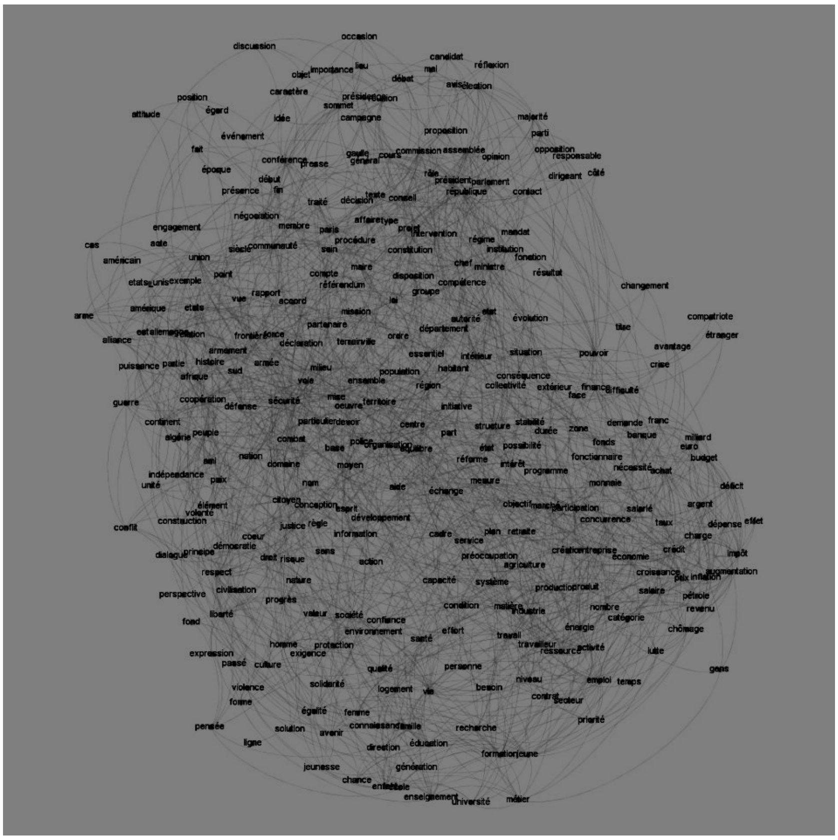
A partir daí, foi possível descrever (i) a dimensão emocional do discurso (*patrie*/pátria→ *amour*/amor → *malheur*/infelicidade, *souffrance*/sofrimento, *haïr*/odiar, *sang*/sangue, etc.), (ii) a dimensão histórica do discurso (*patrie*/pátria → *histoire*/história → *France*/França, *culture*/cultura, *grandeur*/grandeza, *nation*/nação, etc.), (iii) a dimensão política (*patrie*/pátria → *ordre*/ordem→ *autorité*/autoridade, *loi*/lei, *délinquant*/delinquente, *police*/polícia, etc.), (iv) a dimensão espiritual (*patrie*/pátria→ *religion*/ religião → *sentiment*/sentimento, *morale*/moral, *croyance*/crença, etc.) e (v) a dimensão sociológica do discurso (*patrie*/pátria → *famille*/família → *enfant*/criança, *revenu*/renda, *travail*/trabalho, *quartier*/bairro, etc.). Neste último quadro, a estatística volta a encontrar uma espécie de tríptico bem conhecido da história política francesa e completamente inesperado em 2007: *patrie*/pátria, *famille*/família, *travail*/trabalho. Isto sugere – mas trata-se aqui, uma vez mais, de uma simples pista para eventuais pesquisas, de uma forma de leitura, de um filtro hermenêutico – a existência de um discurso ambíguo: ao lado de propostas sociais e de reivindicações que se situam à esquerda, com a evocação de Jean Jaurès ou de Léon Blum, o candidato Nicolas Sarkozy parece dar ao seu discurso uma tonalidade mais neo-pétainista (in)assumida (Mayaffre 2007, 2008 et 2012)*******.

******* Nota do Tradutor: Jean Jaurès, líder socialista, foi assassinado em 1914. A palavra "pétainismo" faz aqui alusão ao Marechal Philippe Pétain, que, em nome da França, assinou o armistício de 22 de junho de 1940, com a Alemanha, e passou a ser, a partir de 10 de julho de 1940, durante a ocupação, o "chefe do Estado francês" (regime de Vichy).

O historiador tem, face ao discurso, muitos imperativos, complexos e complementares, de que não pode fazer abstração, a não ser que subestime a riqueza do material textual: extrair deste material o seu conteúdo, captar-lhe o sentido, compreender o seu funcionamento, medir-lhe a função. Porque, para o historiador, o discurso não é óbvio; ele é ao mesmo tempo a fonte e o objeto, não somente palavras mas acontecimentos, não somente uma imagem do mundo mas uma ação sobre o mundo. A ciência histórica – do arquivo, que se encontra no montante, até a escrita histórica, que se encontra na jusante – é *narratio*, isto é, ao mesmo tempo *narrativa*, *representação* e *configuração* do mundo através da linguagem.

O trabalho sobre as coocorrências, graças ao qual elaboramos esta contribuição, é sem dúvida a forma mais finalizada dos tratamentos digitais ou logométricos. Particularmente, o sentido de uma palavra só pode ser definido pelos seus contextos de utilização e, do ponto de vista linguístico, o contexto imediato de uma palavra pode ser então reduzido, na sua mais simples expressão, à soma das suas coocorrências (Mayaffre 2014). Globalmente, sobretudo, o texto pode ser concebido, no que se refere à sua arquitetura profunda, como um entrelaçamento de palavras que estão em correlação umas com as outras, numa rede de relações entre formas linguísticas ou um espaço reticular no qual as palavras agem como se fossem ecos e *tecem* – segundo a etimologia – o *texto* (ilustração 9).

Figura 9. Visão do texto como espaço lexical reticular (*corpus* presidencial 1958-2014) (de Gephi, versão 8.0).



Este tratamento simboliza, assim, a complementaridade da leitura humana, intuitiva e linear, com uma leitura que se beneficia de ferramentas e que é reticular. Com o tratamento estatístico das ocorrências (cálculo de especificidades, distância intertextual, riqueza e variação do vocabulário, etc.), que os historiadores aprenderam a praticar dos anos 1980 para cá, o tratamento estatístico das coocorrências somente põe em destaque o fato de as ciências humanas (de que a história é parte) se terem transformado hoje em *ciências humanas digitais*: a mudança técnica do suporte do arquivo, da tablete de argila para a tablete digital, passando pelo papel e pela revolução de Gutenberg, representa um compromisso fundamental para o futuro da disciplina, para a nossa visão do texto, para a nossa maneira de abordar o sentido, para os nossos atos de leitura – isto é, para os nossos protocolos interpretativos.

Em verdade, algumas tantas décadas depois de Marc Bloch, Lucien Febvre (Febvre 1953), Alphonse Dupront (Dupront 1969, 1970), Régine Robin (Robin 1973), Antoine Prost (Prost 1988) ou, mais recentemente, de Jacques Guilhaumou (Guilhaumou *et al*. 1994; Guilhaumou 2006), é só hoje, através da digitalização sistemática do arquivo universal, e através da leitura digital deste arquivo, que o historiador vem a ser novamente questionado sobre a sua relação

com os fatos linguísticos. Sempre cultivada pelos grandes espíritos, a ideia de uma história total só se pode bem caracterizar pela relação que têm os historiadores com as ciências da linguagem, que o digital, na sua totalidade, revisita imperiosamente; porque, ontem como hoje, não pode haver história sem arquivo (Web), sem *corpus* (digital) e sem linguagem; como não há linguagem sem história.

REFERÊNCIAS

Jean-Michel Adam, *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Armand Colin, 2008.

Jean-Paul Benzécri, *L'analyse des données*, Paris/Bruxelas/Montreal, Dunod, 1973, 2 vol.

Sergio Bolasco, *L'analisi automatica dei testi. Fare ricerca con il text mining*, Roma, Carocci, 2013.

Étienne Brunet, *Ce qui compte*, Paris, Champion, 2011.

Louis-Jean Calvet et Jean Véronis, *Les mots de Nicolas Sarkozy*, Paris, Seuil, 2008.

Corpus, « La distance intertextuelle », n° 2 (coordenação de Xuan Luong), 2003 (http://corpus.revues.org/).

Robert Darnton, *Apologie du livre*, Paris, Gallimard, 2011.

Alphonse Dupront, « Sémantique historique et histoire », *Cahiers de lexicologie*, I-II, 1969, p. 15-30.

Alphonse Dupront, « Langage et histoire », *Actes du XIIIe Congrès International des Sciences Historiques*, Moscou 16-23 août 1970, t. 1, Naouka, Moscou, 1970-1973, p. 186-254.

Lucien Febvre, *Combats pour l'histoire*, Paris, Colin, 1953.

Jack Goody, *Pouvoirs et savoirs de l'écrit*, Paris, La Dispute, 2007.

Jacques Guilhaumou, « L'histoire du discours et la lexicométrie », *Histoire & Mesure*, n° 3/4, 1986, p. 27-46.

Jacques Guilhaumou, *Discours et événement. L'histoire langagière des concepts*, Besançon, PU de Franche-Comté, 2006.

Jacques Guilhaumou, Denise Maldidier et Régine Robin, *Discours et archive. Expérimentations en analyse du discours*, Liège, Mardaga, 1994.

Serge Heiden et Pierre Lafon, « Cooccurrences. La CFDT de 1973 à 1992 », *Des mots en liberté, Mélanges Maurice Tournier*, Paris, ENS Éditions, 1998, t. 1, p. 65-83.

Dominique Labbé, *Le vocabulaire de François Mitterrand*, Paris, P-FNSP, 1990.

Dominique Labbé et Denis Monière, Le *discours gouvernemental. Canada, Québec, France (1945-2000)*, Paris, Champion, 2003.

Langages, « Unité(s) du texte », n° 163 (coordenação de D. Legallois), 2006.

Ludovic Lebart et André Salem, *Statistique textuelle*, Paris, Dunod, 1994.

Claire Lemercier & Claire Zalc, *Méthodes quantitatives pour l'historien*, Paris, Repères, 2008.

Dominique Longrée & Sylvie Mellet, « Syntactical Motifs and Textual Structures. Considerations based on the Study of a Latin historical Corpus », dans "New Approaches in Text Linguistics", *Belgian Journal of Linguistics*, n° 23, 2009, p. 161-174

Xuan Luong, Dominique Longrée, Sylvie Mellet et Jean-Pierre Barthélemy, « Représentations du texte pour la classification arborée et l'analyse automatique de corpus. Application à un corpus d'historiens latins », Revista *Mathématiques et sciences humaines,* n° 187, 2009, p. 107-121.

Pascal Marchand, *Le grand oral. Le discours de politique générale de la V^e^ République*, De Boeck, Bruxelas, 2007.

Damon Mayaffre, *Le poids des mots. Le discours de gauche et de droite dans l'entre-deux-guerres. Maurice Thorez, Léon Blum, Pierre-Étienne Flandin et André Tardieu 1928-1939*, Paris, Champion, 2000.

Damon Mayaffre et Xuan Luong, « Arbres et généalogie politique. Représentation arborée du discours de Jacques Chirac (1995-2002) », *Histoire et Mesure*, vol. 18, n° 3/4, 2003, p. 289-311.

Damon Mayaffre, *Paroles de président. Jacques Chirac et le discours présidentiel sous la V^e^ République*, Paris, Champion, 2004.

Damon Mayaffre, « Vocabulaire et discours électoral de Sarkozy: entre modernité et pétainisme », *La Pensée*, n° 352, oct.-déc. 2007, p. 65-80.

Damon Mayaffre, « Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence », *JADT 2008*, Lyon, PU de Lyon, vol. 2, p. 811-822.

Damon Mayaffre, *Mesure et démesure du discours. Nicolas Sarkozy (2007-2012)*, Paris, P-ScPo, 2012.

Damon Mayaffre, « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », *JADT 2014*, Paris, Inalco-Sorbonne nouvelle, p. 15-32.

Michèle Monte et Gilles Philippe (dir.), *Genres et textes: déterminations, évolutions, confrontations*, Lyon, PU de Lyon, 2014.

Charles Muller, *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977.

Bénédicte Pincemin *et alii*, « Concordanciers: Thème et variations », *JADT 06*, Besançon, PuFC, p. 773-784.

Antoine Prost, *Vocabulaire des proclamations électorales de 1881, 1885 et 1889*, Paris, PUF, 1974.

Antoine Prost, « Les mots », in René Rémond (dir.), *Pour une histoire politique*, Paris, Seuil, 1988, p. 255-286.

François Rastier, *Arts et sciences du texte*, Paris, PUF, 2001.

François Rastier, *La mesure et le grain. Sémantique de corpus*, Paris, Champion, 2011.

Régine Robin, *Histoire et Linguistique*, Paris, Armand Colin, 1973.

Maurice Tournier, *Propos d'étymologie sociale 2. Des mots en politique*, Lyon, ENS éditions, 2004.

Jean-Marie Viprey, « Philologie numérique et herméneutique intégrative », dans Jean-Michel Adam et Ute Heidmann (éds.), *Sciences du texte et analyse de discours*, Genebra, Slatkine, 2005, p. 51-68.

Recebido: 17 de julho de 2017
Aceito: 19 de julho de 2017