TEXTO DIGITAL
Revista de Literatura, Linguística e Artes

Coocorrências específicas e representações gráficas: o novo "tema" do programa Hyperbase[1]

Specific coocurrences and graphical representations: Hyperbase's new "theme"

Laurent Vanni[a]; Adiel Mittmann[b]

a Universidade Nice Sophia Antipolis, França - laurent.vanni@gmail.com
b Universidade Federal de Santa Catarina, Santa Catarina, Brasil - adiel@mittmann.net.br

Palavras-chave:
Policoocorrência.
Visualização.
Hyperbase.

Resumo: O cálculo dos coocorrentes específicos de uma palavra é um dos métodos estatísticos mais populares na Análise de Dados Textuais. O programa Hyperbase introduziu esse conceito com a função "Tema", que descreve o léxico sobreutilizado na vizinhança de uma forma, de uma estrutura gramatical ou de um lema. Representar essa análise de forma gráfica, porém, é um desafio, pois, mais do que uma simples lista dos coocorrentes específicos, é preciso indicar também a ordem, o desvio e as relações entre cada par de palavras. Com a chegada da versão Web do Hyperbase, propomos agora uma nova abordagem para essa função. O cálculo fica mais aprofundado e passa a identificar a coocorrência de segundo nível. Já a representação gráfica, aposta na simplificação da leitura do resultado da análise e na explicação mais precisa dos cálculos subjacentes.

Keywords:
Polycooccurence.
Visualization.
Hyperbase.

Abstract: Finding word cooccurrences and calculating the specificity scores is one of the most popular statistical methods in the analysis of textual data. Within Hyperbase, there is a "theme" feature for this purpose, which is capable of locating words that are used more commonly near a given word form, grammatical structure or lemma. The graphical representation of such an analysis is often challenging; more than a list of the cooccurring words, it should be able to indicate the order, the score and the relations between pairs of words. Now that Hyperbase has a Web version, this article proposes a new approach for the "theme" feature: the calculation of cooccurrences has been extended to include the second level. The accompanying graphical representation is betting on new visual features in order to simplify reading the results and render the underlying calculation more explicit.

[1] Esta é uma tradução do artigo *Cooccurrences spécifiques et représentations graphiques, le nouveau « Thème » d'Hyperbase*, publicado em 2016 na 13ª edição das *Journées internationales d'Analyse statistique des Données Textuelles*.

1 INTRODUÇÃO

A evolução das ferramentas informáticas nos últimos 15 anos vem contribuindo para moldar a Análise de Dados Textuais (ADT). A *tokenização* de textos com vários milhões de palavras é hoje executada em poucos segundos por processadores extremamente velozes. O volume dos dados não ocupa senão uma pequena parcela dos nossos discos de vários *terabytes*. Os desafios da ADT seguem esta evolução: o tamanho dos dados não é mais uma limitação e a complexidade dos algoritmos recursivos de cálculos estatísticos não representa mais um freio — ao contrário, os algoritmos nos oferecem novos campos de exploração. As visualizações gráficas já vão além da mera ilustração e estimulam uma análise mais profunda: elas utilizam exibições de redes com milhares ou até mesmo milhões de ligações dinâmicas que simultaneamente desconstroem e reconstroem o texto para dele extrairmos diferentes caminhos interpretativos. Essa forma alternativa de leitura, não linear, esclarece o sentido de cada palavra em seu co(n)texto.

A análise das coocorrências específicas é um dos métodos que melhor se presta ao jogo das representações gráficas do texto (VANNI et al., 2014). O que há de mais natural que uma rede para mostrar as ligações ou as atrações entre as palavras? Mais que uma simples lista de coocorrentes, convém exibir a topologia do *corpus* como um todo para avaliar os efeitos do co(n)texto de uma palavra. Os métodos gráficos permitem, assim, representar vários critérios de coocorrência simultaneamente. Se uma linha indica a presença de uma relação entre duas palavras, a espessura da linha pode caracterizar melhor a relação indicando o índice de especificidade.

A partir do cálculo de coocorrência específica já implementado por Étienne Brunet no Hyperbase e de sua função "Tema", detalharemos aqui como a versão Web do programa propõe a passagem da coocorrência simples à policoocorrência (MARTINEZ, 2012). Apresentaremos, em seguida, uma representação gráfica original que permite explorar essas coocorrências de forma dinâmica e aprofundar a abordagem coocorrencial. Finalmente, para ilustrar os resultados obtidos com essa nova ferramenta, utilizaremos um *corpus* produzido a partir dos anais da conferência JADT (Journées internationales d'Analyse statistique des Données Textuelles) de 2000 a 2014. Veremos, através da policoocorrência disponível no Hyperbase, como a evolução da informática e a aparição de novos métodos influenciaram a ADT e as nossas contribuições científicas.

2 ESPECIFICIDADES E COOCORRÊNCIAS

2.1 UM CÁLCULO BEM ESTABELECIDO

O cálculo das especificidades implementado na maior parte dos programas de ADT, baseia-se no modelo hipergeométrico (LAFON, 1980), que utiliza 4 medidas fundamentais para uma dada palavra num *corpus*:

- T = tamanho do *corpus*,
- t = tamanho do texto,
- f = frequência da palavra no *corpus*,
- k = frequência da palavra no texto.

A fórmula que utiliza cálculos de fatoriais e produz a probabilidade de uma palavra ter frequência k no texto é

$$P(x = k) = \frac{f!\,(T-f)!\,t!\,(T-t)!}{k!\,(f-k)!\,(t-k)!\,(T-f-t+k)!\,T!}.$$

Convertido em desvio reduzido[2], esse cálculo caracteriza a sobreutilização ou subutilização de uma palavra com relação à norma (uso médio) em um texto. É aqui que o conceito de texto foi estendido para que fosse adaptado à coocorrência. O texto não é mais um conjunto de parágrafos que constituem uma parte do *corpus*: com a coocorrência, ele se torna o conjunto de passagens que contêm a palavra de pesquisa.

Concatenadas, as passagens[3] pertinentes a uma palavra formam um texto. Para cada palavra, temos também as quatro medidas necessárias ao cálculo hipergeométrico: o tamanho T do *corpus*, o tamanho t do texto (o número de palavras vizinhas à palavra de pesquisa no *corpus*), a frequência f da palavra no *corpus* como um todo e a frequência k da palavra na vizinhança da palavra de pesquisa.

[2] Também conhecido como escore z ou escore padronizado.

[3] O termo "passagem" aqui se refere, por padrão, ao parágrafo, delimitado no arquivo texto pelo fim da linha, mas é possível configurar o sistema para que um determinado número de palavras, antes e depois da palavra de pesquisa, sejam consideradas.

2.2 IMPLEMENTAÇÕES E PARÂMETROS

Por padrão, esse cálculo é efetuado sobre as formas gráficas[4] do texto. Com o Hyperbase Web, aproveitamos a análise morfossintática produzida pelo programa para estender a gama de possibilidades. O usuário pode, assim, analisar a coocorrência com base não apenas nas formas mas também nos lemas e nas categorias gramaticais. Esta camada introduzida pela lematização e pela etiquetagem do texto pode ser aplicada a vários níveis. É possível fixar o tipo da palavra de pesquisa independentemente daquele dos coocorrentes analisados; deixamos o usuário livre para cruzar essas diferentes propriedades linguísticas. Uma palavra de pesquisa pode, portanto, tornar-se um código gramatical e seus coocorrentes, lemas. As possibilidades são numerosas e permitem um ajuste fino da pesquisa de coocorrência.

Com essa nova implementação da função "Tema" no Hyperbase, introduzimos também outros parâmetros que o usuário pode configurar. A filtragem das palavras de acordo com a sua categoria gramatical é um exemplo: o usuário pode solicitar que o *corpus* seja reduzido a uma categoria gramatical específica, como substantivos, adjetivos ou verbos. Ao associar várias categorias, pode-se também extrair um vocabulário maior e demandar, por exemplo, o conjunto de substantivos presentes num texto. A partir de então, são modificados os valores T e t e, consequentemente, os resultados dos cálculos de especificidade. Essa filtragem por categoria gramatical permite tornar o texto mais leve e reduzir o número de palavras a serem analisadas conservando a noção de contexto e de linearidade do texto.

As fronteiras dessa coocorrência são fixadas, por padrão, numa janela definida pelo parágrafo. Esse tamanho de contexto pode ser modificado à vontade para passar de um conjunto grande de várias dezenas de palavras na vizinhança da palavra de pesquisa a um conjunto muito mais restrito de apenas algumas palavras, o que favorece a coocorrência sintática. A interface do Hyperbase Web permite, portanto, que o usuário configure o número exato de palavras antes e depois da palavra de pesquisa.

Esses diferentes parâmetros estão disponíveis numa tela de opções associadas a cada uma das funções do Hyperbase Web. Fixados a valores padrão, eles permitem compreender a coocorrência de forma simples e intuitiva. Para um usuário avançado, contudo, os parâmetros

[4] A forma gráfica é a palavra exatamente como ela se encontra escrita no texto.

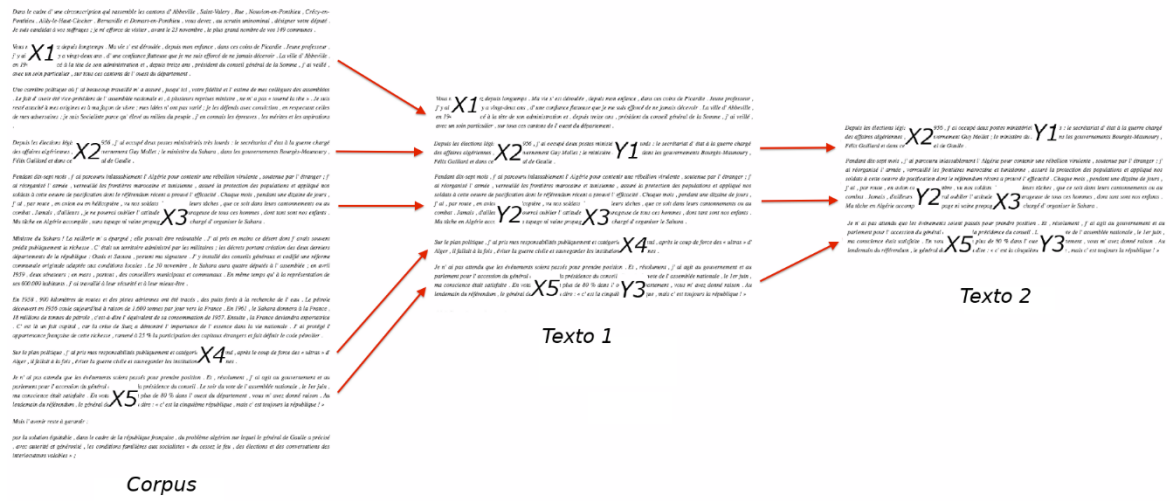
abrem a possibilidade de investigar hipóteses de trabalho mais complexas e específicas. De posse dessas novas opções, estendemos também a função "Tema" para ir além da coocorrência simples: no Hyperbase Web, estamos interessados também na policoocorrência (MARTINEZ, 2003, 2012).

2.3 POLICOOCORRÊNCIA ESPECÍFICA

Convém observar que, na literatura, a policoocorrência (LEBLANC; MARTINEZ, 2006, MARTINEZ, 2012) tem um parente próximo: a coocorrência de segunda ordem ou de ordem superior (GREFENSTETTE, 1994). A ideia das coocorrências de ordem superior é tomar cada coocorrente como ponto de partida (palavra de pesquisa) de um novo cálculo de coocorrência sobre o conjunto do *corpus*. Elas foram aplicadas, por exemplo, à similaridade semântica (LEMAIRE; DENHIÈRE, 2006) e à análise da monossemia (BERTELS; GEERAERTS, 2012). O Hyperbase adota a policoocorrência (MARTINEZ, 2003, 2012), que corresponde à ideia original da função "Tema", a qual visa a explorar a semântica de uma palavra de pesquisa através de seus coocorrentes. O contexto, portanto, é sempre ligado à palavra de pesquisa escolhida inicialmente.

Para chegar à policoocorrência, o cálculo hipergeométrico (Seção 2.1) é estendido através de uma chamada recursiva. Cada um dos coocorrentes é utilizado para definir um novo contexto mais restrito com base na mesma palavra de pesquisa. É extraído do *corpus* um novo texto, produzido a partir de todas as passagens que contêm simultaneamente a palavra de pesquisa e o coocorrente selecionado (Figura 1). Nesse novo contexto, um índice de coocorrência de segundo nível é então calculado para cada uma das palavras presentes. Assim, todo coocorrente da palavra de pesquisa recebe sua própria lista de palavras específicas. Essas novas ligações de especificidade entre as palavras formam a policoocorrência no Hyperbase.

Figura 1. Seleção do contexto da palavra de pesquisa X e subsequentemente do contexto de Y, coocorrente de X.



Esta camada de policoocorrência torna o resultado complexo o suficiente para que uma olhadela seja incapaz de avaliá-lo. Quando a função "Tema" é selecionada no Hyperbase Web, uma lista das palavras nos é apresentada juntamente com um índice de especificidade para cada uma. A coocorrência de segundo nível não pode ser apresentada dessa mesma maneira; as combinações resultantes são demasiadamente volumosas para serem estruturadas numa lista simples. É aqui que as visualizações gráficas dinâmicas podem revelar-se preciosas para avaliar a rede de ligações coocorrenciais e destacar diferentes caminhos interpretativos com um simples sobrevoo do *mouse*.

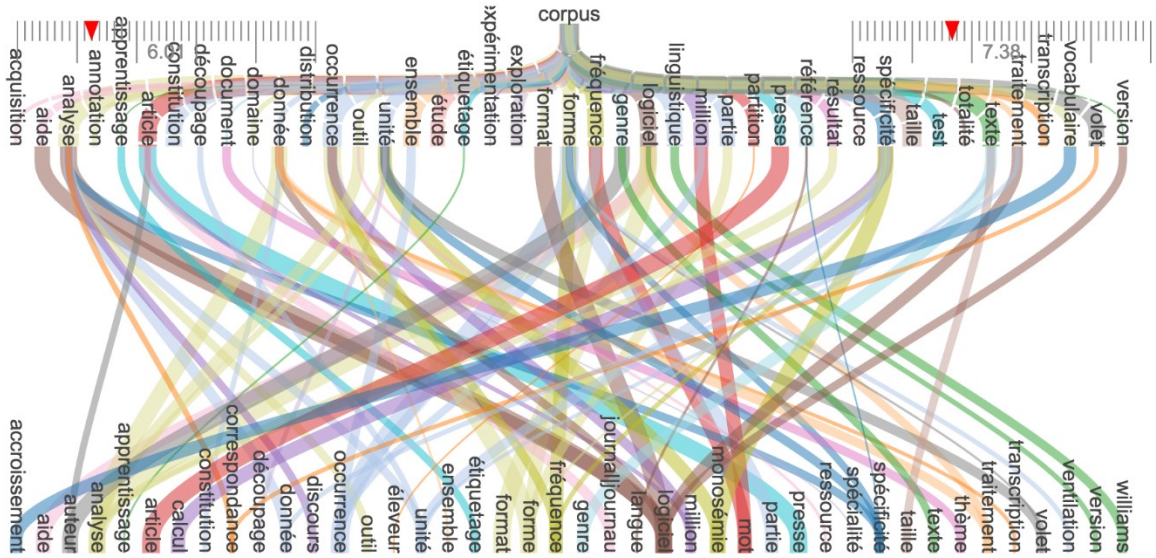
3 REPRESENTAÇÕES GRÁFICAS

A proposta de visualização que aqui fazemos visa a mostrar a coocorrência e a policoocorrência de forma conjunta, como mostra a Figura 2. Diferentemente de Martinez (2012), esta visualização é capaz de tratar simultaneamente um grande número de palavras. Para aproveitar a capacidade de processamento disponível nos navegadores contemporâneos, a biblioteca D3.js (BOSTOCK et al., 2011) foi utilizada. A D3.js fornece ferramentas modernas em JavaScript que respondem às necessidades da visualização gráfica *online*. Essa biblioteca é fornecida com exemplos prontos para serem modificados[5] e com comandos básicos que permite a criação de nossas próprias visualizações. Até o momento, não existiam

[5] Este método já foi utilizado no Hyperbase para exibir os grafos de especificidade e a distribuição das palavras no *corpus*.

implementações com a D3.js que permitissem a exibição de um gráfico que respondesse às necessidades da policoocorrência. Criamos, portanto, uma visualização completamente nova com essa biblioteca.

Figura 2. Policoocorrência: visualização em camadas dinâmicas.

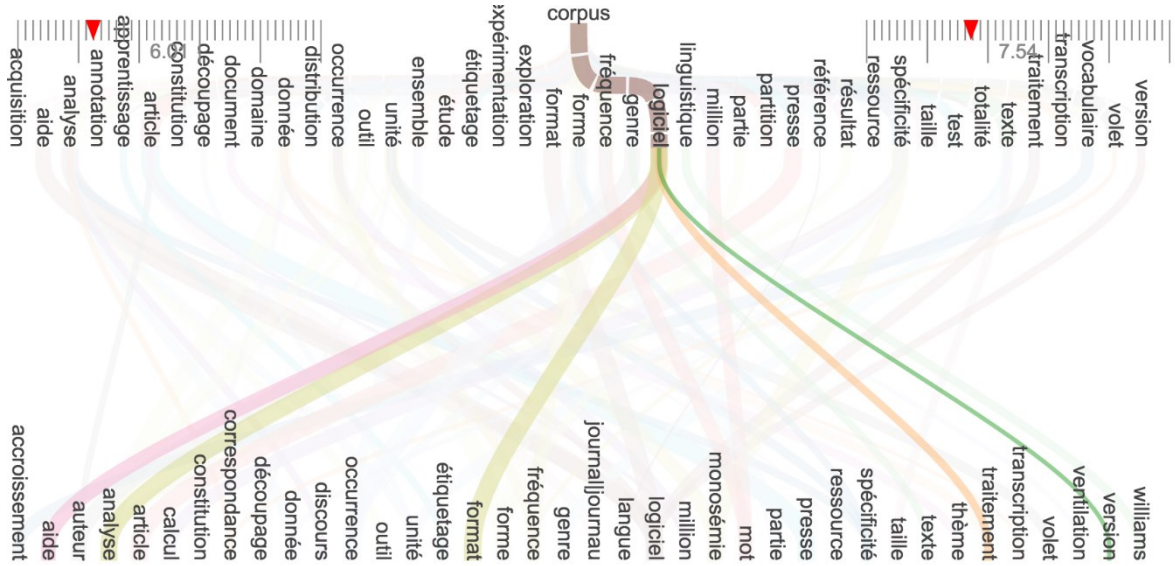


Optamos por uma visualização em camadas sucessivas, que mostra a palavra de pesquisa na cabeça do gráfico, a coocorrência simples na primeira camada e a policoocorrência na segunda. As palavras do primeiro e do segundo nível são filtradas em função de limiares de especificidade independentes. Quando os limiares são ajustados[6], as alterações são aplicadas à visualização através de uma animação fluida que inicia imediatamente.

Em cada camada, as palavras são ordenadas alfabeticamente e dispostas da esquerda para a direita. Se não há espaço suficiente para as exibir horizontalmente, elas são, então, viradas em 90 graus para a direita; se, mesmo assim, falta espaço, o tamanho da fonte é reduzido suficientemente para que as palavras sejam exibidas sem sobreposição.

[6] Os limiares de especificidade podem ser ajustados graficamente através da movimentação do *mouse* sobre uma régua.

Figura 3. Filtragem do gráfico com o sobrevoo do *mouse*.



As linhas que ligam as palavras representam a especificidade da palavra da camada inferior com relação à palavra da camada superior. Essas linhas são, portanto, orientadas, mas uma flecha não é exibida por questões de legibilidade. O gráfico é lido naturalmente de cima para baixo.

Se a palavra de pesquisa escolhida pelo usuário é muito frequente e os limiares selecionados são baixos, a quantidade de linhas na visualização pode tornar-se visualmente fatigante. Para que o usuário possa explorar as informações disponíveis confortavelmente, criamos um sistema de filtragem dinâmica, ativado pelo sobrevoo do *mouse*. Esse sistema destaca as associações de uma palavra em meio a centenas de outras, apagando as ligações inúteis. Um exemplo é dado na Figura 3. As palavras são dispostas de maneira sequencial para permitir que as associações sejam varridas em sucessão rápida, o que possibilita a identificação de resultados pertinentes na massa de dados exibidos pelo gráfico.

Ao selecionar, com o *mouse*, uma palavra da primeira camada e outra da segunda, a visualização adota um modo secundário de exibição (Figura 4). Aqui, a palavra escolhida na primeira camada e a palavra de pesquisa são colocadas lado a lado no topo do gráfico. A primeira camada passa a ser destinada aos coocorrentes especificamente ligados às palavras selecionadas. A segunda camada fica reservada à segunda palavra selecionada. O objetivo aqui é evidenciar todas as ligações específicas que existem entre a palavra de pesquisa e dois de seus coocorrentes, ocupando todo o espaço disponível. O gráfico propõe ao usuário,

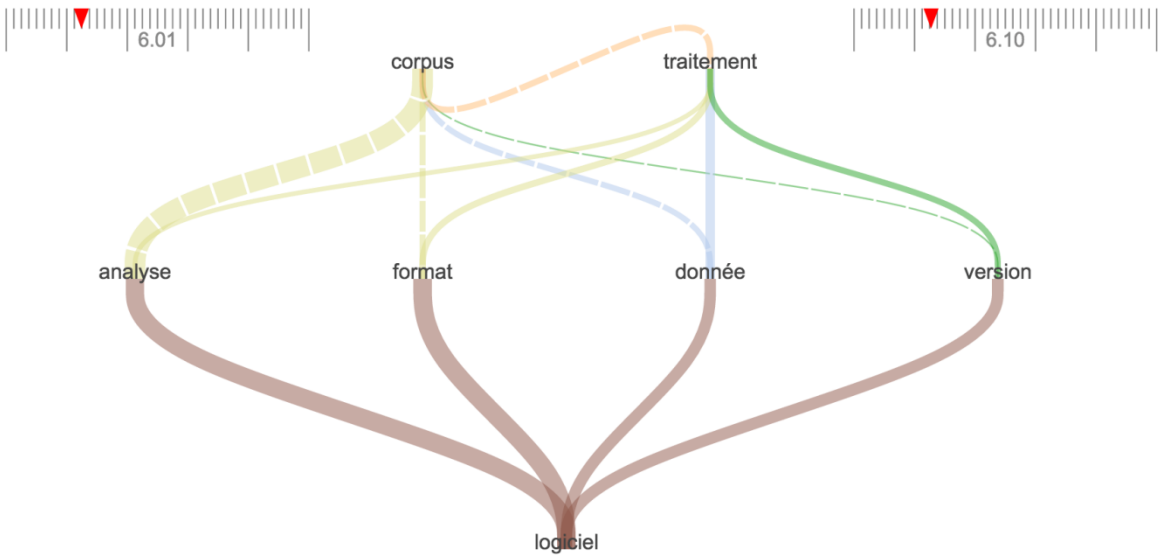
portanto, vários caminhos interpretativos que passam pelo intermédio de outras palavras específicas do ambiente de palavras selecionadas. Para evitar confusão, as linhas de coocorrência direta (que partem diretamente da palavra de pesquisa) são exibidas de maneira pontilhada, enquanto aquelas oriundas da policoocorrência ficam sólidas.

Nesta visualização, o desvio reduzido[7] é representado pela espessura das linhas. Essa espessura *p*, em *pixels*, é calculada em função de cada desvio reduzido *e* através da fórmula

$$p = m\ln(1 + e - \theta_i) + k,$$

em que *m* representa um fator de escala e θ_i é o limiar de especificidade da camada superior. O parâmetro *k* representa a espessura mínima que resulta quando $e = \theta_i$. Essa fórmula garante que a espessura é informativa e visualmente adequada ao longo de uma ampla gama de desvios reduzidos e de limiares.

Figura 4. *Zoom* no ambiente lexical das palavras "tratamento" (traitement) e "programa" (logiciel) em torno da palavra de pesquisa "corpus".



Ao fixar limiares diferentes para a primeira e segunda camadas, o usuário pode produzir uma visualização significativa e informativa. Entretanto, a fórmula que leva em consideração o limiar de especificidade gera espessuras diferentes para os dois conjuntos de linhas: dadas

[7] Cabe lembrar que o índice de especificidade apresentado na Seção 2.1 é convertido, no Hyperbase, em desvio reduzido.

uma linha da palavra de pesquisa à primeira camada e outra linha da primeira camada à segunda, uma espessura idêntica não significa que as palavras-alvo têm o mesmo desvio reduzido. Essa espessura deve ser comparada somente em relação às ligações da mesma camada. O valor exato do desvio reduzido é exibido dinamicamente quando o *mouse* passa por cima da linha. As cores das linhas são arbitrárias e selecionadas dentre uma lista de cores predefinidas. Uma função de *hash*[8] é aplicada à palavra-alvo para determinar a cor da ligação escolhida. Esse método distribui as cores de forma a aumentar o contraste e a legibilidade do gráfico.

Essa forma de visualização da coocorrência e da policoocorrência nos permite explorar nossos *corpora* sob um novo ângulo. Enquanto a coocorrência simples para na ligação direta entre duas palavras, o gráfico aqui apresentado nos permite ir mais longe e identificar diferenças de "Tema" que somente aparecem à luz da policoocorrência. Assim, em duas partes distintas do mesmo *corpus*, mesmo quando constatamos os mesmos coocorrentes para a palavra de pesquisa, podemos diferenciar o valor semântico ao analisar os policoocorrentes associados. O exemplo que segue apresenta um caso assim, encontrado no *corpus* dos anais da JADT.

4 CASO DE USO

Vimos na Seção 3 que as representações gráficas são ferramentas indispensáveis que revelam todas as ramificações de um resultado estatístico, como é o caso da coocorrência específica. Mas a ADT em geral tomou de fato esse rumo nos últimos anos? Ela apostou na visualização gráfica como forma de representar dados cada vez mais complexos? Para saber se as práticas de ADT se alteraram ao longo dos últimos anos, tomamos o conjunto dos artigos publicados nas JADTs (Journées internationales d'Analyse Statistique des Données Textuelles) entre 2000 e 2014, o que corresponde a 14 anos e a um total de 2,5 milhões de palavras.

Para esta análise diacrônica, dividimos o *corpus* em dois períodos. O primeiro, entre 2000 e 2006, está associado ao aumento na capacidade das ferramentas informáticas e à aparição de novos meios de comunicação. O segundo, entre 2008 e 2014, foi marcado pelo acesso aos

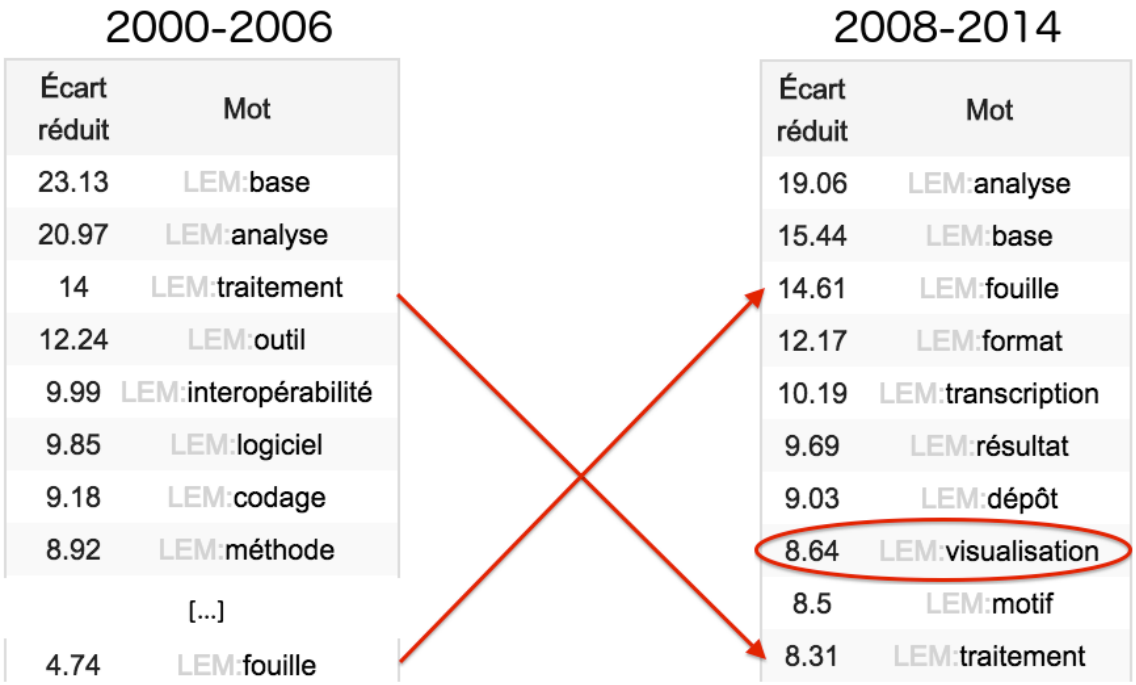
[8] Uma função de *hash*, em informática, calcula uma espécie de impressão digital dos dados. Ela garante uma repartição homogênea dos valores retornados.

recursos textuais através da utilização da internet — a Web forneceu um volume de informações nunca antes atingido.

No início dos anos 2000, o cálculo das coocorrências específicas do lema “dado” (donnée), mostrado na Figura 5, nos ensina que esses “dados” estão fortemente associados ao lema “tratamento” (traitement, +14). O tratamento já é automático ou semiautomático e assistido por “programas” (logiciels, +9,85). Ele apoia-se também sobre os “métodos” (méthodes**,** +8,92) tradicionais da ADT. Estamos diante de uma abordagem tradicional em que a qualidade do *corpus* é mais importante que a quantidade. As ferramentas estatísticas servem sobretudo para testar uma hipótese de trabalho. Ainda estamos longe de uma abordagem heurística. A “busca” (fouille) de dados é uma prática incipiente e ainda pouco presente. Esse termo faz uma aparição tímida (+4,74) na lista de coocorrências, que mostra um interesse nascente por esse tipo de análise que necessita um grande volume de dados.

Entre os anos 2008 e 2014, o tamanho dos *corpora* tratados aumentou em dez vezes e a fronteira da “busca” (fouille) de dados foi superada (+14,61). Ela vence até mesmo o simples “tratamento” (traitement, 8,31), com um escore que posiciona essa abordagem entre os três primeiros coocorrentes mais específicos do lema “dado” (donnée). Com a democratização da internet e o livre acesso ao recurso textual, os *corpora* adquiriram vários milhões de palavras. Eles são compostos de novas formas de dados, oriundos de diversas fontes como jornais eletrônicos, blogues, tweets e redes sociais.

Figura 5. Fragmento das especificidades do lema “dado” (donnée) nas duas partes do *corpus*. A coluna “Écart réduit” corresponde aos desvios reduzidos e a coluna “Mot” às palavras.

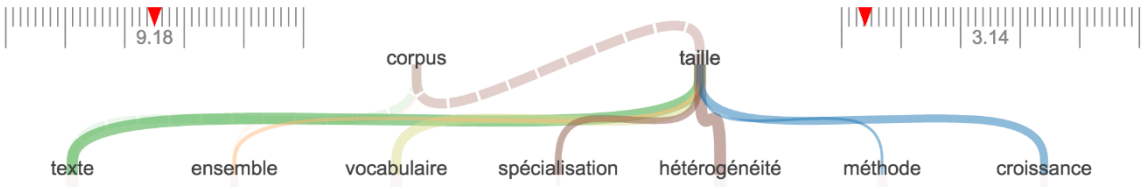


Como a comunidade de ADT se propõe a representar esses novos recursos? Observamos que o lema “visualização” (visualisation, +8,64) é agora um dos 10 coocorrentes mais específicos da lista. Essa mudança metodológica fez com que até mesmo o uso da palavra “método” (méthode) recuasse na lista dos coocorrentes diretos da palavra “dado” (donnée). Para entender essa evolução, é necessário avançar nossa análise com outros conceitos muito presentes nas publicações das JADTs, a saber, o “corpus” e o “texto” (texte). Vamos agora contemplar a policoocorrência na vizinhança dessas palavras para extrair alguns elementos de resultado.

Tanto do ponto de vista do “texto” (texte) ou do *corpus*, uma questão recorrente na ADT é o “tamanho” (taille), ou seja, o número de palavras que se deseja analisar. Com um escore de +11,38, em 2000–2006, e +8,63, em 2008–2014, em torno da palavra de pesquisa “corpus”, gostaríamos de observar a policoocorrência para distinguir as diferenças entre esses dois períodos. As Figuras 6 e 7 nos mostram, respectivamente para o primeiro e segundo períodos, a policoocorrência da palavra “tamanho” (taille). O resultado parece comprovar a diferença metodológica que já era pressentida na lista de coocorrentes da Figura 5. O tamanho entre os anos 2000–2006 (Figura 6) era um eco da “heterogeneidade” (hétérogénéité) do *corpus*. A

passagem a uma escala maior, com um grande volume de dados, era assunto de discussão quando da composição do *corpus*. O problema da heterogeneidade, que introduz um viés interpretativo, era, portanto, estreitamente ligada à questão do tamanho. Com a Figura 7, constatamos que os novos recursos digitais oriundos da Web nos anos 2008–2014 reconfiguraram as problemáticas e as questões ligadas à ADT.

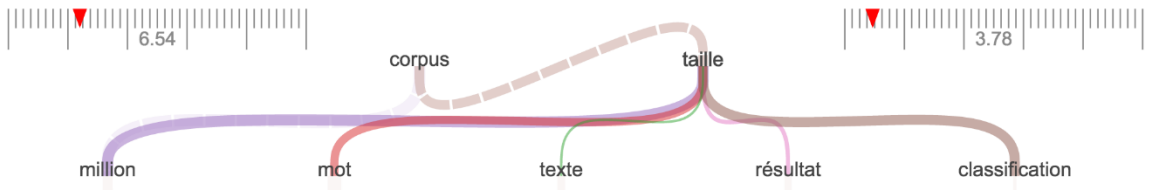
Figura 6. Policoocorrência da palavra "tamanho" (taille), no ambiente da palavra de pesquisa "corpus" — período entre 2000 e 2006.



Entre 2008 e 2014, o problema da heterogeneidade cede lugar aos "milhões" (millions) de "palavras" (mots), como mostra a Figura 7. Na era digital, os recursos textuais *online* possibilitam que os *corpora* atinjam tamanhos incomensuráveis. Já em 2012, Étienne Brunet (2012) propunha estudar 44 bilhões de palavras (através do *Google Books*) com a sua ferramenta Hyperbase. Depois, em 2014, propúnhamos revisitar esse *corpus* faraônico, agora com 86 bilhões de palavras (BRUNET; VANNI, 2014). Certamente havíamos criticado a heterogeneidade necessariamente associada a esse volume de dados e os erros de etiquetagem morfossintática produzidos por uma anotação rápida. Mas fomos forçados a constatar que já era possível analisar a evolução da grafia das palavras do francês nos últimos 200 anos por meio das nossas ferramentas estatísticas habituais.

Essa mudança de prioridade também se observa no gráfico pela presença do lema "resultado" (résultat), que figura no contexto de "corpus" e de "tamanho" (taille). Se considerarmos que esses resultados estão enriquecidos com gráficos, como nos leva a crer o lema "visualizações" (visualisations), que é coocorrente de "dado" (donnée), temos ali um meio de explorar esses novos *corpora*. A heterogeneidade torna-se aqui uma questão passível de ser mensurada graficamente. A título de exemplo, podemos citar um outro coocorrente que também aparece, o lema "classificação" (classification). As classificações dos textos são derivadas de grafos, como as árvores, que utilizam métodos estatísticos para nos dar informações preciosas acerca da constituição do *corpus*.

Figura 7. Policoocorrência da palavra “tamanho” (taille) no ambiente da palavra de pesquisa “corpus” — período entre 2008 e 2014.



Essa análise contrastiva ilustra a evolução das práticas de ADT. Os métodos são sensíveis aos meios tecnológicos. A digitalização dos dados era, por si só, um desafio há 16 anos; hoje, as humanidades digitais nos dão a esperança de *corpora* sem limites de tamanho. As ferramentas devem levar em conta esse novo paradigma. As visualizações gráficas que surgiram nesses últimos anos nos dão agora um meio de explorar em detalhe esses *corpora*. As escolhas ergonômicas a serem feitas são numerosas e não seria surpreendente se nos próximos anos aparecessem novas problemáticas ligadas a essas escolhas.

5 5. CONCLUSÃO

Entre a atualização do cálculo e a renovação da visualização coocorrencial, o Hyperbase Web nos oferece atualmente uma nova função “Tema”. Essa função do programa permite que os usuários analisem seus *corpora online* com o auxílio da policoocorrência. A análise do co(n)texto de uma palavra é levada a cabo através de visualizações gráficas dinâmicas executadas diretamente no navegador do usuário. O Hyperbase segue, portanto, a evolução dos meios e dos métodos tecnológicos e propõe um outro ponto de vista sobre a coocorrência.

O conjunto de parâmetros disponíveis na ferramenta permite adaptar os dados às necessidades do usuário. A filtragem por códigos morfossintáticos e a lematização fornecem elementos para hipóteses bastante sofisticadas que se baseiam na coocorrência, sem contudo distanciarem-se do cálculo original que garantiu o sucesso dessa função. A rede coocorrencial obtida pela policoocorrência foi objeto de um estudo aprofundado da representação gráfica. O usuário dispõe atualmente de um conjunto de ferramentas que lhe permite percorrer diferentes caminhos interpretativos em poucos cliques.

O Hyperbase Web está, portanto, entre as novas tecnologias das quais dispõe a ADT. A técnica e os meios evoluíram desde o início dos anos 2000. Os *corpora* mudaram no que se

refere à sua natureza e ao seu tamanho, como refletem os anais das JADTs, que são facilmente encontrados em formato digital a partir do ano 2000. Essa base está disponível livre e gratuitamente[9] — uma outra tendência observável entre as ferramentas atuais.

REFERÊNCIAS

BERTELS, A.; GEERAERTS, D. L'importance du recoupement des cooccurrents de deuxième ordre pour étudier la corrélation entre la spécificité et la monosémie. In: 11èmes Journées internationales d'Analyse statistique des Données Textuelles, p. 135–147, 2012.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, v. 17, n. 12, p. 2301–2309, dez. 2011.

BRUNET, E. Au fond du GOOFRE, un gisement de 44 milliards de mots. In: 11èmes Journées internationales d'Analyse statistique des Données Textuelles p. 7–21, 2012.

BRUNET, E.; VANNI, L. GOOFRE version 2. In: 12èmes Journées internationales d'Analyse statistique des Données Textuelles, p. 105–119, 2014.

GREFENSTETTE, G. Corpus-derived first, second and third-order word affinities. In: Proceedings of EURALEX, p. 279–290, 1994.

LAFON, P. Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, ENS Editions, v. 1, n. 1, p. 127–165, 1980.

LEBLANC, J.-M.; MARTINEZ, W. L'analyse contrastive des réseaux de cooccurrence: Le monde dans les discours des présidents de la Cinquième République. In: 8èmes Journées internationales d'Analyse statistique des Données Textuelles, p. 603–615, 2006.

LEMAIRE, B.; DENHIÈRE, G. Effects of High-Order Co-occurrences on Word Semantic Similarities. *Current Psychology Letters - Behaviour, Brain and Cognition*, v. 18, n. 1, p. 1, fev. 2006.

MARTINEZ, W. Au-delà de la cooccurrence binaire... Poly-cooccurrences et trames de cooccurrence. *Corpus*, n. 11, 2012.

MARTINEZ, W. *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. 2003. Tese (Doutorado) – Université de la Sorbonne nouvelle - Paris 3.

[9] Para acessar o *corpus*, basta carregar a base "jadt" no endereço <http://hyperbase.unice.fr/>.

VANNI, L.; LUONG, X.; MAYAFFRE, D. Arbre et co-occurrences. Nouvel outil logométrique sur le net. Application au discours de François Hollande. In: 12èmes Journées internationales d'Analyse statistique des Données Textuelles, p. 639–649, 2014.

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