TEXTO DIGITAL
Revista de Literatura, Linguística e Artes

As Novidades da Versão 10 do Programa HYPERBASE*

Quelques nouveautés dans HYPERBASE version 10

Etienne Brunet[a]

a Universidade de Nice Sophia Antipolis, França - prof.etiennebrunet@gmail.com

Palavras-chave:
Hyperbase. Corpus.

Resumo: Trata-se aqui fundamentalmente de uma apresentação dos novos recursos do programa Hyperbase, na sua versão mais recente (10). Trata-se consequentemente, para o autor, de um exercício de explicação dos novos recursos metodológicos agora utilizados, ou funções. Este artigo está, em consequência, mais particularmente destinado às pessoas que já conhecem e/ou utilizam o programa Hyperbase e estão familiarizadas com os modelos matemáticos e recursos informáticos aqui mencionados. Os demais poderão encontrar algumas dificuldades de leitura, em função basicamente da terminologia. O leitor vai, por outro lado, encontrar aqui uma série de referências a outros programas e métodos, particularmente TXM, BOOTSTRAP, ALCESTE e IRAMUTEQ, LEXICO e GEPHI.

Keywords:
Hyperbase. Corpus.

Abstract: Il s'agit fondamentalement, dans cet article, de présenter la nouvelle version du logiciel Hyperbase (version 10). Il s'agit en conséquence, pour l'auteur, d'un exercice d"explication des nouvelles ressources méthodologiques ou fonctions mises en oeuvre. Dès lors, cet article est plus particulièrement destiné aux personnes qui ont une connaissance préalable du logiciel Hyperbase et sont déjà familiarisées avec les modèles mathématiques et les ressources informatiques qui y sont mentionnés. Les autres pourront y trouver certaines difficultés de lecture, en raison notamment de la terminologie utilisée. Le lecteur trouvera par ailleurs ici des références à d'autres programmes et méthodes, notamment TXM, BOOTSTRAP, ALCESTE et IRAMUTEQ, LEXICO et GEPHI.

* Tradução de Carlos Alberto Antunes Maciel – BLC (UMR 7320 / NuPILL). Título original, em francês: **Quelques nouveautés dans HYPERBASE version 10.** Nota do tradutor: Os exemplos apresentados são provenientes, no essencial, de um *corpus* francês (discursos do General De Gaulle); assim, as palavras em francês, apresentadas nos diferentes exemplos, são sistematicamente traduzidas (uma só vez, no entanto). Quando se trata de funções (AJOUTER – ACRESCENTAR, por exemplo, ou ainda CORPUS EXTÉRIEURS), o tradutor adotou o mesmo sistema, mantendo no entanto as maiúsculas. As palavras, bastante numerosas, que aparecem em algumas figuras, não foram traduzidas, exceção feita daquelas que, devidamente integradas às figuras, são mencionadas no próprio texto, fora consequentemente das figuras.

1 – O programa HYPERBASE[**] **pode ser agora baixado gratuitamente** (a partir do sítio http://logometrie.unice.fr, primeira página), onde havia, e onde ainda atualmente há, cerca de sessenta bases de dados, literárias, linguísticas ou políticas, também disponibilizadas gratuitamente. Acrescenta-se a isto, por outro lado, o arquivo HYPERBASE.EXE, em formato *Zip*, com o qual o programa pode ser instalado (e com o qual novas bases de dados podem ser geradas). A instalação pode ser feita em qualquer computador equipado com Windows (sistemas XP, Vista, 7, 8 e 10). Quando foi criado, em 1989, o preço de um programa parecia ainda ter uma relação com o seu real valor ou a sua utilidade; depois disso, no entanto, os ventos fortes da *open source* e do *2.0* tudo varreram e acabaram com tais preconceitos. Como nenhum salário está vinculado à criação e ao desenvolvimento do produto, é bom que estejamos enfim livres do peso das obrigações comerciais e bom também que possamos seguir exemplos como o da equipe TXM[***] ou ainda da IRAMUTEq[****], assim como o de tantos outros empreendimentos universitários.

2 – A segunda novidade que merece destaque é a **simplificação**. Tirando a versão de base, que não lematiza os dados e que não é praticamente mais utilizada, uma versão única é agora proposta ao usuário, qualquer que seja a língua dos dados e qualquer que seja o lematizador utilizado. O utente deve simplesmente indicar a língua dos dados (francês, inglês, espanhol, alemão, português ou italiano) e o lematizador que utiliza – TreeTagger[*****] ou Cordial[******]. Uma única versão, chamada HYPERnew.tbk, substitui a partir de agora todas as precedentes variantes Hypercor, Hypertag, Hyperang, Hyperger, etc... Não se deve no entanto confundir a

[**] Nota do tradutor: Etienne Brunet, autor do programa Hyperbase, é Professor da Universidade de Nice (e hoje aposentado). Criou e foi o primeiro Diretor da antiga URL 9 (Unidade de Pesquisa Linguística 9), do CNRS (*Centre National de Recherche Scientifique* ou Centro Nacional de Pesquisa Científica – Universidade de Nice). A antiga URL 9 muito evoluiu desde a sua criação e, sumariamente, corresponde hoje ao Laboratório "*Bases, Corpus et Langage*" (ou Bases, Corpus e Linguagem – BCL) – UMR 7320 (*Unité Mixte de Recherche* ou Unidade Mixta de Pesquisa), do CNRS – Universidade de Nice. O programa Hyperbase foi concebido na fase inicial, ainda na URL 9, e muito evoluiu nos últimos anos, sempre com Etienne Brunet, no âmbito das atividades científicas da UMR 7320 – CNRS.

[***] Nota do tradutor: Textométrie. Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte (Textometria. Unidade que federa as pesquisas e as evoluções tecnológicas em textometria para a criação de uma plataforma de programas informáticos abertos – <http://textometrie.ens-lyon.fr/spip.php?rubrique96>).

[****] Nota do tradutor: IRAMUTEQ = *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* (Interface de R para as Análises Multidimensionais de Textos e de Questionários – <http://www.iramuteq.org/>).

[*****] Nota do tradutor: *Centre de traitement automatique du langage* (Centro de processamento automáticos da linguagem – <http://cental.fltr.ucl.ac.be/treetagger/>).

[******] Nota do tradutor: *ATALA – Association pour le traitement automatique des langues (ATALA* – Associação para o processamento automático das línguas – <https://www.atala.org/CORDIAL-Universites-ou-Cordial>).

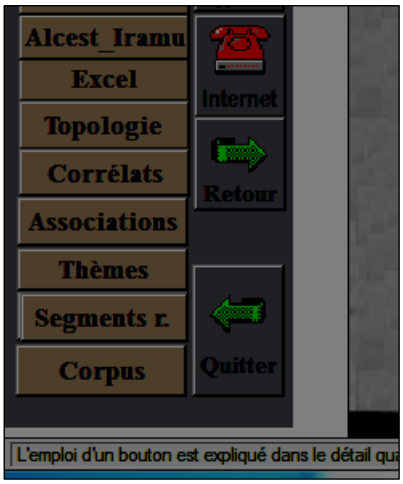
língua dos dados com a língua do programa. O diálogo com o usuário faz-se, em todos os casos, em francês. Uma versão do programa "fala", no entanto, português, graças à tradução feita por Carlos Maciel[*******]. Uma versão em língua inglesa já está prevista, mas não foi ainda disponibilizada.

3 – O número de textos disponíveis para todo e qualquer tipo de processamento aumentou muito ultimamente, e as ambições de cada utente quando da constituição de um *corpus* cresceram nas mesmas proporções. Ora, a arquitetura do programa impunha limites, tanto no que se refere ao número de textos (75) quanto no que diz respeito à extensão do *corpus* (sempre menos de 10 milhões de palavras). Esta estrutura foi modificada e a capacidade do programa foi **multiplicada por dois**.

1 A FUNÇÃO CORPUS EXTÉRIEURS (CORPORA EXTERNOS)

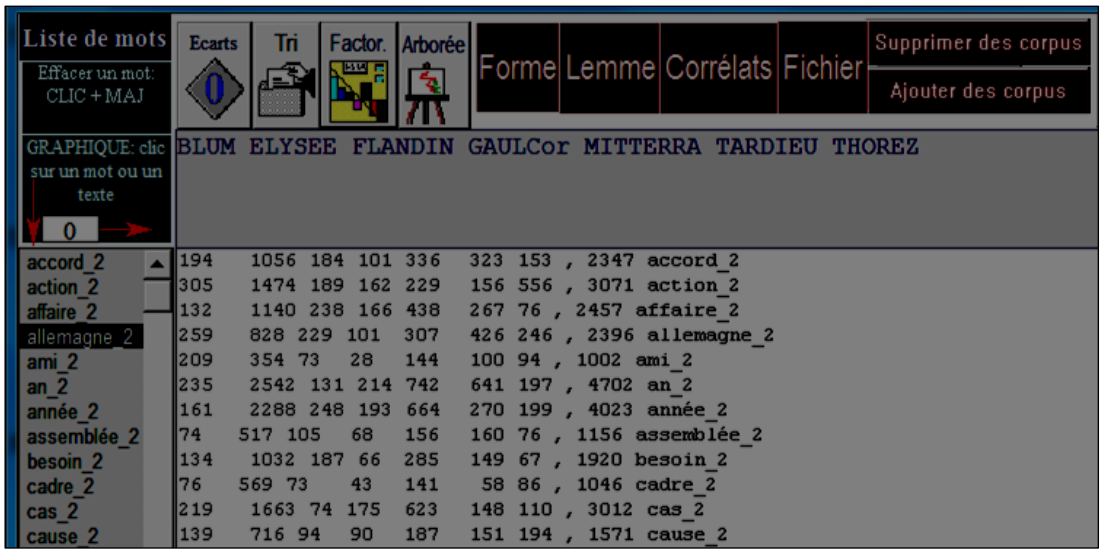
A proposta, neste caso, sem sair do *corpus* que está sendo analisado, é de fazer uma comparação com todo e qualquer outro *corpus* já anteriormente constituído e analisado pelo programa HYPERBASE. O último item, ***Corpus***, situado na margem esquerda do menu principal, permite a execução desta função.

Figura 1. Localização da função no menu principal.



[*******] Nota do tradutor: Trata-se, neste caso, da versão 9, bastante utilizada no Brasil, particularmente no NuPILL – UFSC (Núcleo de Pesquisa em Informática, Literatura e Linguística – Universidade Federal de Santa Catarina, em Portugal, particularmente no Centro de Linguística da Universidade Nova de Lisboa (Profa. Teresa Rijo da Fonseca Lino), e em outros países de língua portuguesa.

Figura 2. A página da função Corpus. Escolha das palavras,
de todo e qualquer corpus, e dos processamentos.



A primeira operação consiste em fazer a **seleção do *corpus*** (com uma ou mais unidades), solicitando, segundo o caso, os botões AJOUTER (ACRESCENTAR) ou SUPPRIMER (SUPRIMIR) CORPUS. Não é para isto necessário que estas bases estejam localizadas na pasta C:\HYPERBAS. Elas podem ter sido gravadas em outro lugar, num DVD, numa caneta USB ou num disco externo. Não são com efeito estas bases que são propriamente exploradas, mas dois arquivos que estão associados a elas e que têm o mesmo nome da base, um com o sufixo .TXT (que é o dicionário de frequências das grafias) e o outro com o sufixo .TX2 (que é o dicionário de lemas).

A segunda operação consiste em **escolher as palavras** que vão formar as linhas de um quadro em que as colunas representam cada *corpus* selecionado. É preciso então distinguir entre as grafias e os lemas, já que os arquivos em que se encontram uns e outros são diferentes (esta distinção impõe-se no momento em que é feita a escolha do *corpus*). Quando uma lista de palavras for criada, é possível exportá-la para que ela possa ser importada mais tarde, se o usuário quiser a ela voltar (botão FICHIER – ARQUIVO). Por outro lado, uma lista do *corpus* que está sendo analisado, presente na página LISTE (LISTA), onde são feitas todas as escolhas a fazer no âmbito do programa HYPERBASE, pode ser transplantada e aumentada para o estudo de cada *corpus*. O botão CORPUS EXTÉRIEURS (CORPORA EXTERNOS) permite fazer isso. Finalmente, a lista dos correlatos, que reúne as formas plenas (e mais particularmente os substantivos) mais frequentes do *corpus* pode servir de critério para a seleção das palavras. Ao selecionar o botão CORRELATS, o que primeiramente se faz é a lista de todas as formas

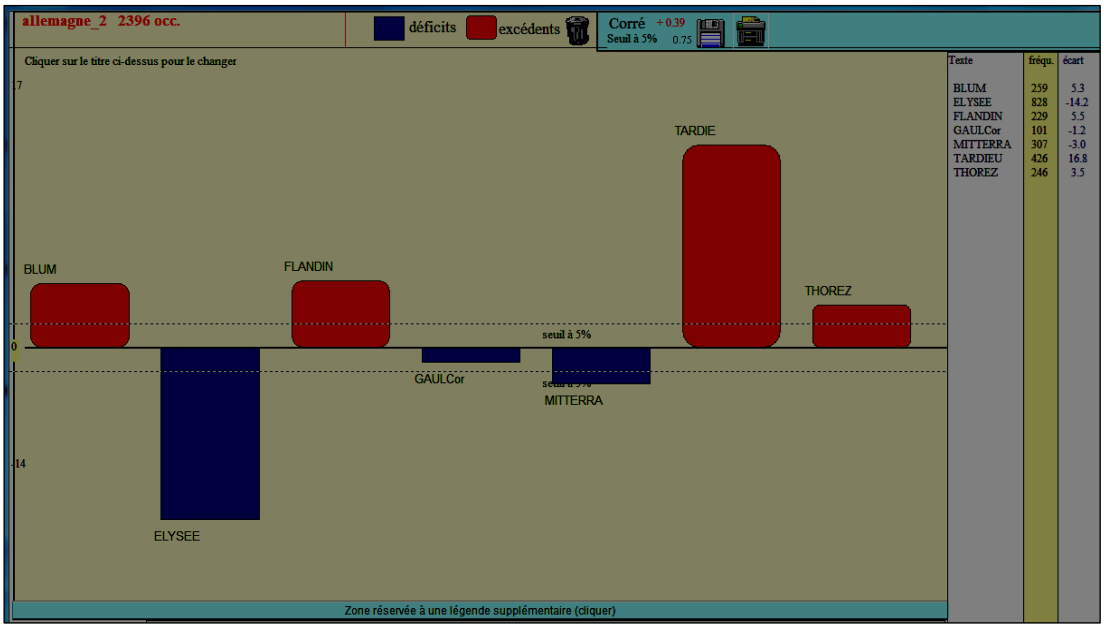
plenas que foram consideradas como correlatos quando um *corpus* é comparado com outro. Esta lista passa depois disso por uma triagem segundo a repartição das palavras no *corpus*, a começar pelos elementos que aparecem em todos os diferentes *corpora*; na parte final, encontramos as palavras que aparecem num *corpus* somente. Sabendo que só 200 palavras, no máximo, podem ser levadas em consideração (para uma mais fácil leitura dos histogramas e das análises de fatores), o usuário tem a possibilidade de escolher as formas que estão na parte superior ou na parte inferior da lista (ou até mesmo na parte central). No primeiro caso, são as palavras que estão mais presentes nos diferentes *corpora* que serão analisadas; no segundo caso, temos as unidades mais especializadas e mais marcantes. O botão CORRELATS supõe que nos diferentes *corpora* considerados – ou em alguns deles pelo menos – a lista dos correlatos, na qual figuram os substantivos mais frequentes, já foi gerada. Se assim não for, o *corpus* em questão será levado em consideração, mesmo se ele não tem participação na escolha das palavras.

A comparação que se faz entre os diferentes *corpora* permite **abrir** muito o campo de exploração dos dados. Considerando que a extensão de um *corpus* analisado graças ao programa HYPERBASE pode dificilmente ultrapassar 20 milhões de palavras, o limite pode chegar perto do bilhão de palavras se juntarmos dezenas de *corpora*. A leitura dos dicionários não é seguramente em tais casos instantânea, mas o tempo necessário para a consulta não é de modo nenhum excessivo: um ou dois segundos são suficientes para ler o dicionário de um *corpus* de um milhão de palavras.

As funções disponíveis são as que se aplicam a um quadro:

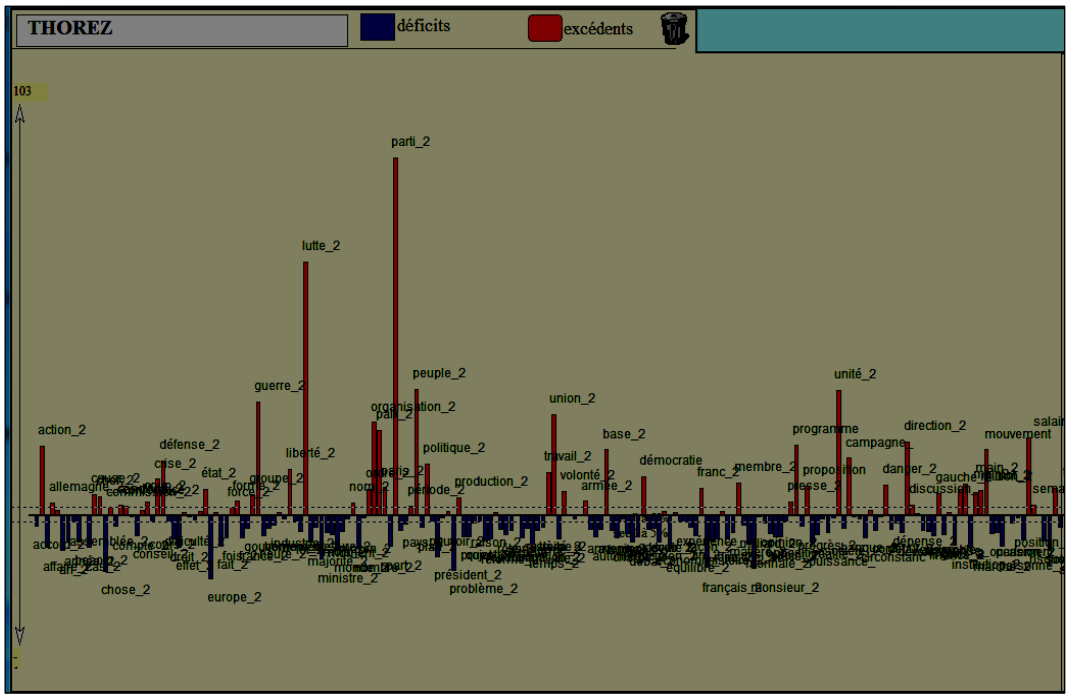
1 – **histograma de uma palavra** nos diferentes *corpora* (basta clicar numa palavra, na margem esquerda). Atenção: um CLIQUE com uma MAIÚSCULA apaga a palavra.

Figura 3. Histograma da palavra Alemanha em sete diferentes corpora políticos.



2 – **histograma de um *corpus*** através das palavras que nele se encontram (basta clicar no nome do *corpus*, na margem superior).

Figura 4. Histograma do corpus Thorez com cerca de cinquenta substantivos



3 – função de **triagem** (alfabética ou hierárquica) aplicada às palavras da lista (a ordem do *corpus* acompanha a sequência da seleção).

4 – **análise de fatores** com base nas frequências absolutas ou no cálculo dos desvios reduzidos.

Figura 5. Análise de fatores dos correlatos melhor distribuídos

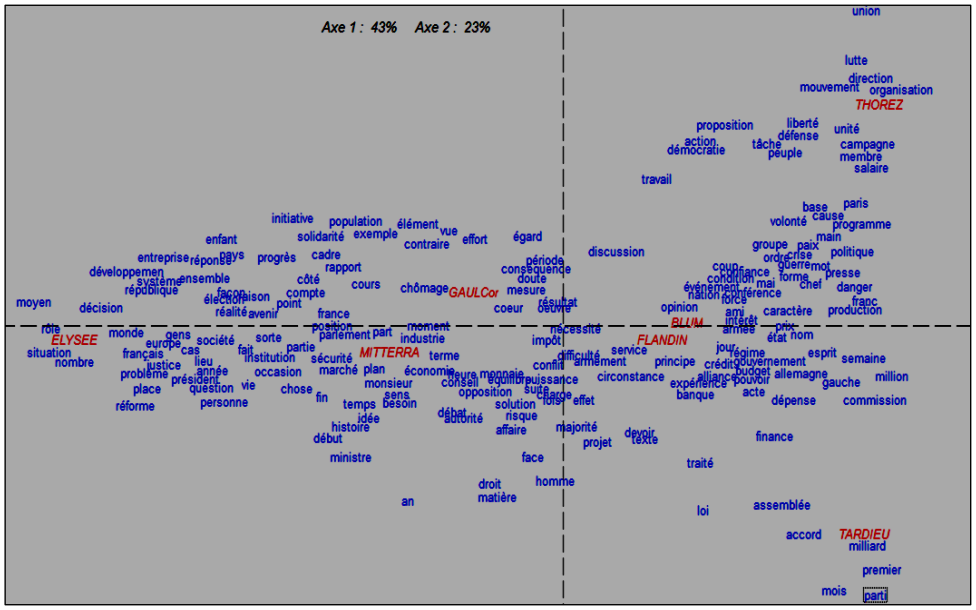
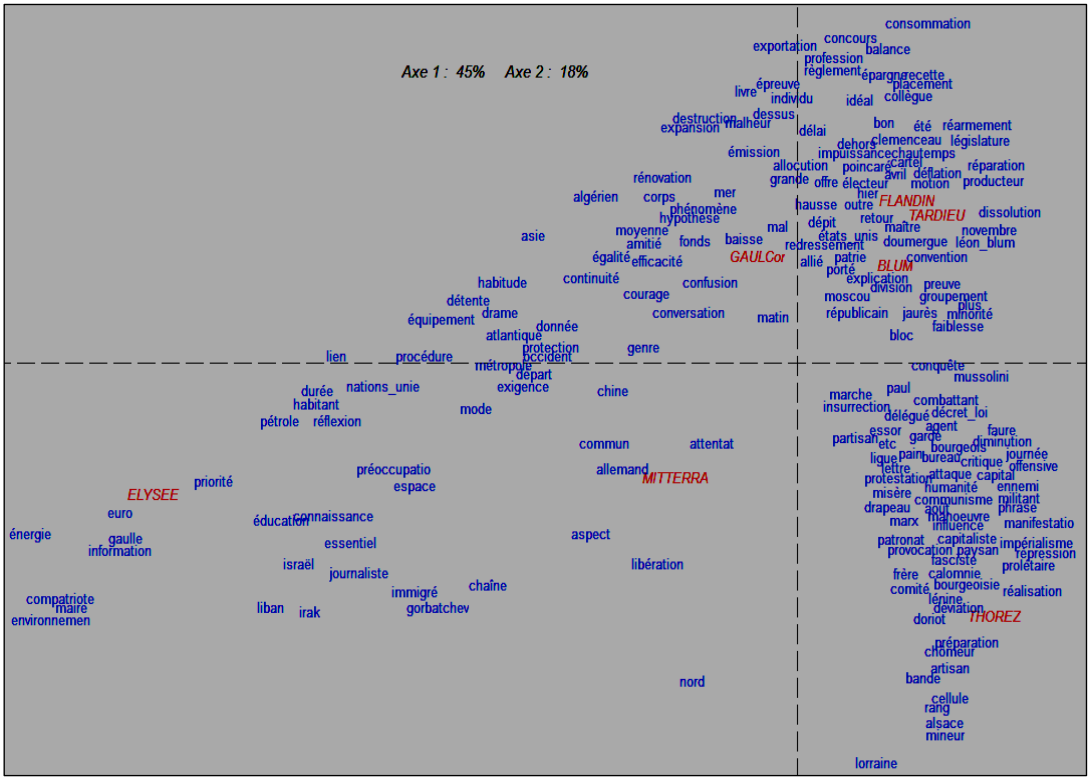


Figura 6. Análise de fatores dos correlatos mais marcantes.



5 – **análise em árvores** gerada a partir dos dados relativos às frequências ou aos desvios.

Figura 7. Análise em árvores das colunas

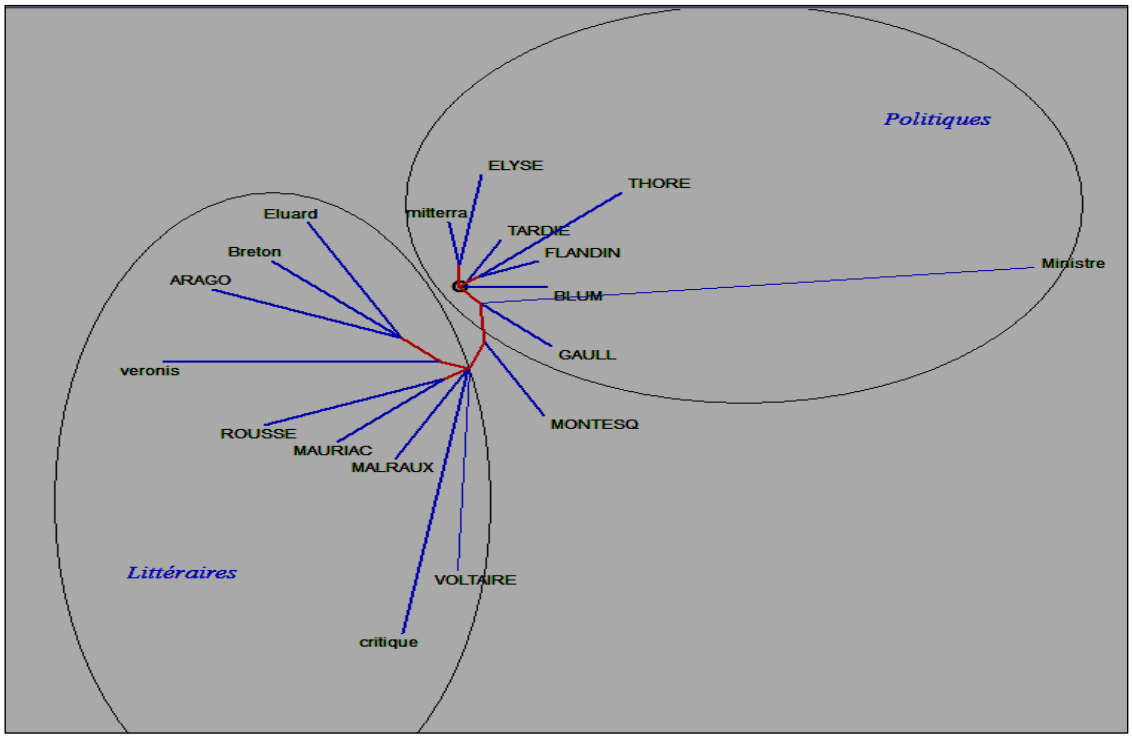
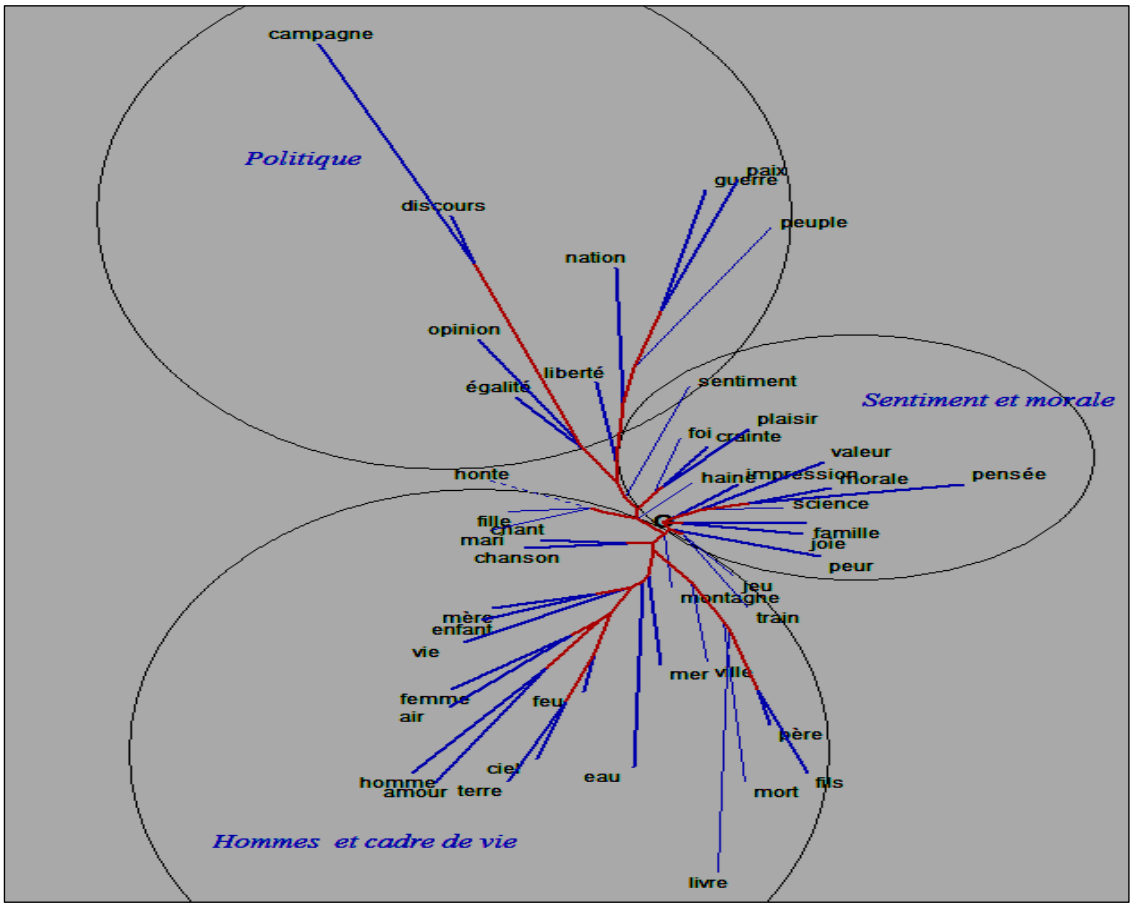


Figura 7 bis. Análise em árvores das linhas



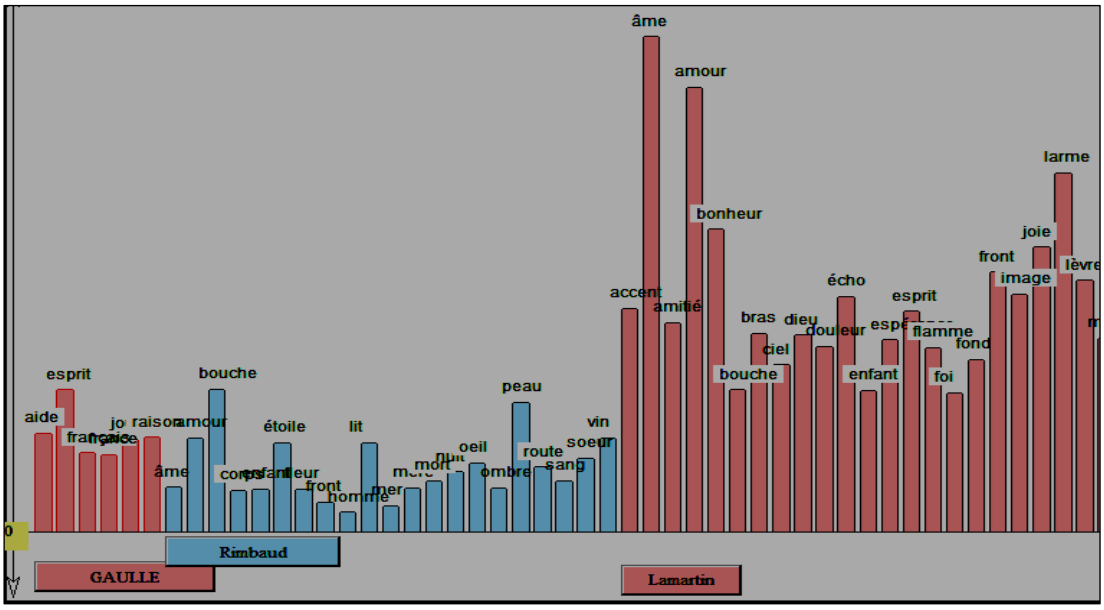
1.1 ALGUMAS OBSERVAÇÕES

1 – Como o volume dos dados é muito importante, torna-se inútil e é até mesmo perigoso apelar para a hipergeometria. E a **lei normal**, que produz os desvios reduzidos, justifica-se plenamente em tais casos. Graças ao botão correspondente, é possível escolher, alternando, entre uma apresentação a partir dos desvios ou das frequências.

2 – Cabe observar que o quadro não é tratado como um quadro de contingências, que seria por si só suficiente. O total da linha corresponde sem dúvida ao total marginal (horizontal), mas o mesmo não acontece com o total da coluna, que resulta do total do *corpus* inteiro, e não do total das palavras selecionadas. E o mesmo acontece com o efetivo total, que resulta da acumulação de todos os *corpora*. O desvio reduzido é exatamente aquele que teria sido obtido se puséssemos os *corpora* todos juntos, na sequência, sem nem mesmo excluir uma palavra. Fazemos, assim, com que sejam implicitamente considerados todos os dados subjacentes, o que faz com que as palavras selecionadas se destaquem, em pano de fundo, com relação a todas aquelas que nem mesmo vemos.

3 – A extensão dos *corpora*, em se tratando do programa Hyperbase, não é uma verdadeira novidade. Este enfoque já tinha sido tentado com as coocorrências. Tínhamos então percebido que as coocorrências são observadas na totalidade do *corpus* e que não é possível cruzar estes dados com as partes deste mesmo *corpus*. A única comparação possível faz-se então em termos e na escala dos *corpora*. Buscando os arquivos onde estão registrados os resultados da pesquisa relativa às cooccorrências, podemos prolongar o **histograma das formas coocorrentes** de uma palavra dada e estender as observações a outros *corpora*, sempre com a mesma palavra. Foi possível, então, ver que a mesma palavra, *cœur* (coração), por exemplo, podia estar associada, ou ter "parceiros" diferentes, segundo os autores analisados e com valores muito divergentes. Nas obras de De Gaulle, as relações que a palavra *cœur* tem com outras são em geral bastante frias, e só têm fundamentalmente interesse quando se trata das unidades *France* (França) e *raison* (razão). Em Rimbaud, o coração é mais carnal e aproxima-se de palavras como *peau* (pele), *corps* (corpo), *amour* (amor), *lit* (cama) e *vin* (vinho). Em Lamartine, o coração ocupa um amplo espaço, que é partilhado com *amour*, *âme* (alma), *sentiment* (sentimento), *bonheur* (felicidade), *larmes* (lágrimas) e *joie* (alegria).

Figura 8. Histograma das formas coocorrentes da palavra cœur (coração) em diferentes corpora



2 A FUNÇÃO EXPLICITATION DES VOISINAGES (EXPLICITAÇÃO DAS VIZINHANÇAS) NAS ANÁLISES DE FATORES

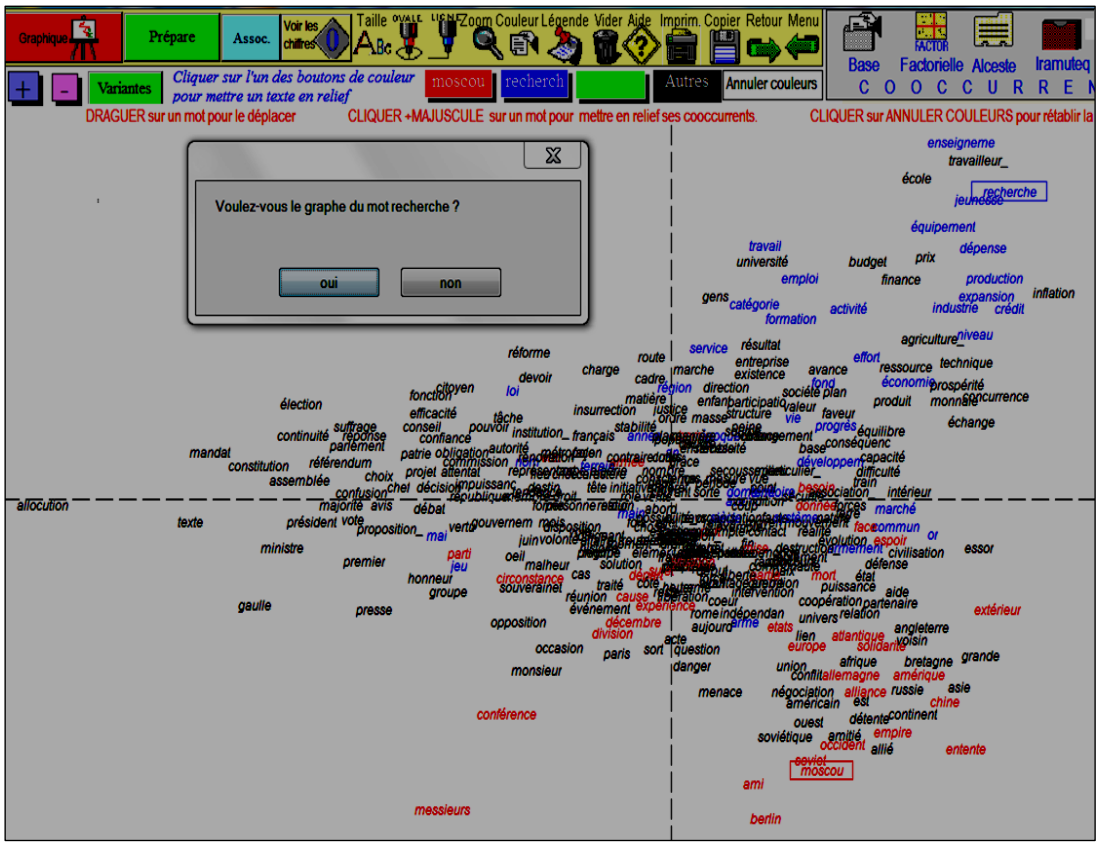
As análises de fatores ou análises de correspondências representam uma armadilha na qual frequentemente caímos. Temos sempre aquela tendência para reunir pontos que aparecem no gráfico, vizinhos uns dos outros, mais ou menos em virtude do princípio segundo o qual "*quem se parece sempre aparece"*. Assim como imaginamos constelações planas, repartidas no espaço plano do firmamento, sem a profundidade que pode fazer com que duas estrelas muito distantes possam se apresentar aos nossos olhos como próximas, pelo menos em aparência, o cruzamento plano de dois fatores pode criar o mesmo tipo de ilusão, já que outros fatores estão escondidos, e tais fatores podem distorcer as vizinhanças, mesmo quando estas vizinhanças são estreitas. Além disso, esta proximidade, no melhor dos casos, reflete uma semelhança de perfil, em que gostos e rejeições são partilhados. Isto não quer necessariamente dizer que há relações reais e individualizadas. As palavras que aparecem lado a lado no gráfico podem nunca se ter encontrado no texto, como acontece com as pessoas que vão a uma passeata e gritam as mesmas palavras de ordem, de mãos dadas, e que nem mesmo se conhecem.

O perigo da hiperinterpretação é ainda mais importante quando se trata de coocorrência. Nos casos mais tradicionais, em que cruzamos as linhas, onde se encontram as palavras, e as colunas, onde se encontram as partes do *corpus*, bem sabemos que o que aproxima as palavras

não é um laço que vincula uma à outra, mas a tendência que as caracteriza e que faz com que estejam presentes nas mesmas partes do *corpus*, mas não necessariamente juntas, nas mesmas passagens, como a Lua e o Sol, que se encontram juntos no céu mas que estão raramente juntos. Quando, no entanto, o cálculo se faz diretamente com relação à copresença das palavras no espaço estreito da frase ou do parágrafo, somos necessariamente levados a ver na vizinhança uma comunidade. E, para lutar contra esta tendência, é preciso calcular o valor real desta vizinhança para separar aquilo que resulta de um verdadeiro espaço partilhado do que resulta do acaso.

Ora, o cálculo hipergeométrico é feito pelo programa HYPERBASE mas para uma palavra somente, considerada então como o polo, com o seu círculo de amigos próximos e distantes desenhado numa figura arborescente em que a espessura dos arcos, o tamanho das letras e a distância com relação ao próprio polo representam a maior ou menor força dos vínculos existentes. Pensamos, então, que era preciso mandar esta informação para a análise de fatores das coocorrências. Um clique numa palavra basta para a designar, e para que ela fique vermelha, por causa de uma espécie de auréola que surge. Esta cor vermelha propaga-se e atinge a vizinhança, mas de forma seletiva. Alguns vizinhos imediatos nem mesmo se mexem e ficam como se fossem blocos de gelo. Mas alguns elementos mais longínquos se manifestam indicando que pertencem ao clã, e alguns deles até mesmo se encontram para além dos eixos da análise. No conjunto, a mesma cor é, no entanto, partilhada na vizinhança mais imediata do polo escolhido. Ao escolher alguns outros pontos, obtemos a imagem de um estádio onde nem os jogadores de um mesmo time nem os torcedores misturam as suas cores.

Figura 9. As verdadeiras relações das palavras escolhidas (de cor azul) e Moscou (de cor vermelha), na análise de fatores da coocorrência.

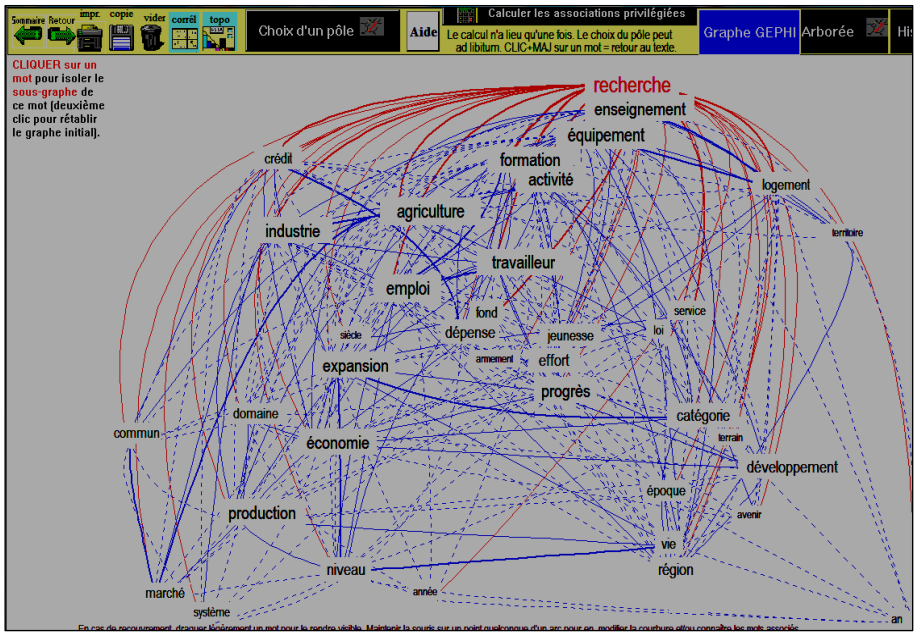


Na ilustração aqui apresentada, relativa aos discursos de De Gaulle, escolhemos dois pontos bem distantes no sentido vertical: a palavra *recherche* (pesquisa), de cor azul, na parte superior do gráfico, e a palavra *Moscou*, na parte inferior, de cor vermelha, como é cabível. Estes dois polos simbolizam a oposição entre a política interna, considerado o seu aspecto econômico, e a política externa (a outra ponta do triângulo, bem à esquerda, diz respeito às instituições e às eleições).

Para confirmar as indicações dadas pelas cores, o usuário é convidado a reunir num mesmo grafo os diferentes partidários declarados e coloridos do campeão selecionado. Eles ficam agora entre si, numa sala reservada para eles, onde os assentos são distribuídos de acordo com a hierarquia. O campeão instala-se bem no alto do estrado, com aqueles com quem tem relações diretas ou episódicas, reunidos aos poucos numa **estrutura que tem a forma de um sino** e onde os laços ao mesmo tempo se aglutinam e se ramificam. Não estamos aqui falando de uma caderneta de endereços em que os nomes não têm relação uns com os outros; estamos falando de um espaço macio e movediço onde um movimento browniano intervém para multiplicar as trocas. A partir do momento em que um tipo de vínculo ultrapassa o limite para

além do qual o acaso pode ser invocado, isto vai estar simbolizado por um traço pontilhado, mais fino ou mais espesso, segundo a importância do tráfego. O detalhe é por vezes instrutivo e indica aquilo a que damos o nome de **artéfacts de la phraséologie** (artefatos da fraseologia). Não deveria haver nenhum tipo de troca entre *marché* (mercado) e *commun* (comum), já que se trata de uma mesma realidade. Da mesma maneira, uma lematização mais rigorosa poderia impedir a desintegração do *niveau de vie* (nível de vida). Tirando estes casos triviais, as relações fortemente marcadas podem ser facilmente interpretadas, como o acordo que há entre *agriculture* (agricultura) e *marché commun* (mercado comum), a relação estreita que há entre *crédit* (crédito) e *dépense* (gastos, despesas), entre *formation* (formação) e *emploi* (emprego), entre *logement* (alojamento) e *équipement* (equipamento), etc... Cabe indicar que estes acordos bilaterais são negociados independentemente do polo, no texto inteiro, o que é diferente da função THEMES (TEMAS), que só leva em consideração as passagens onde o polo aparece.

Figura 10. Grafo da palavra recherche (pesquisa)



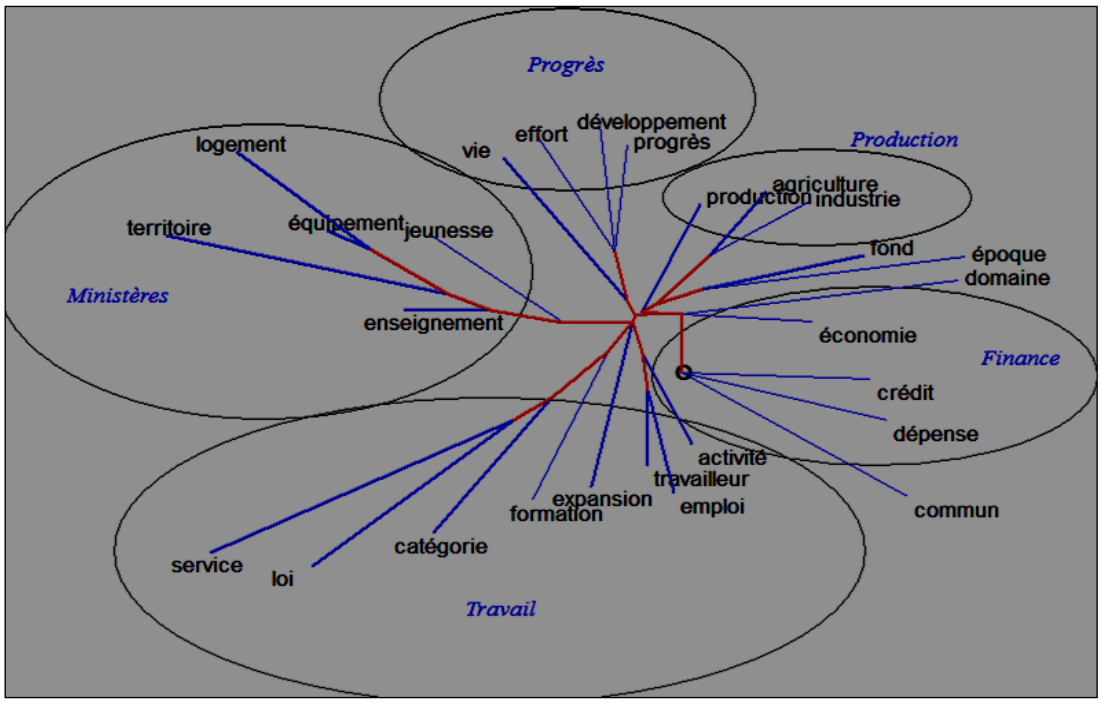
3 A FUNÇÃO *ZOOM* NOS GRAFOS DE COOCORRÊNCIA

É preciso reconhecer que a lisibilidades destes grafos deixa muitas vezes a desejar. É difícil acompanhar com os olhos todos os traços que partem de um nó ou que conduzem a ele. Duas ou três pequenas coisas ajudam, no entanto, a ver tudo isso mais claramente.

1 – Encontramos a primeira no início do procedimento, no diálogo que propõe uma **seleção** severa, média ou branda. Se formos rigorosos na escolha do limite, vamos deixar de lado todo aquele emaranhado inextricável de traços pontilhados para conservar unicamente as relações que são altamente significativas.

2 – Se preferirmos uma solução mais elegante, vamos recorrer à **análise em árvores**[********], fornecendo-lhe, para isso, um quadro das relações que há entre as coocorrências. As palavras são, neste caso, obrigadas a escolher um único campo. E obtemos, então, um mapa hidrológico em que cada ponto só tem um caminho possível para o escoamento das águas. Assim, a bacia da figura abaixo está dividida em 4 ou 5 lotes; a *recherche* está, então, como podemos ver, associada, segundo o caso, ao *travail* (trabalho), à *production* (produção), à *finance* (finanças), ao *progrès* (progresso) ou ao *gouvernement* (governo).

Figura 11. Análise em árvores dos coocorrentes da palavra recherche (pesquisa)

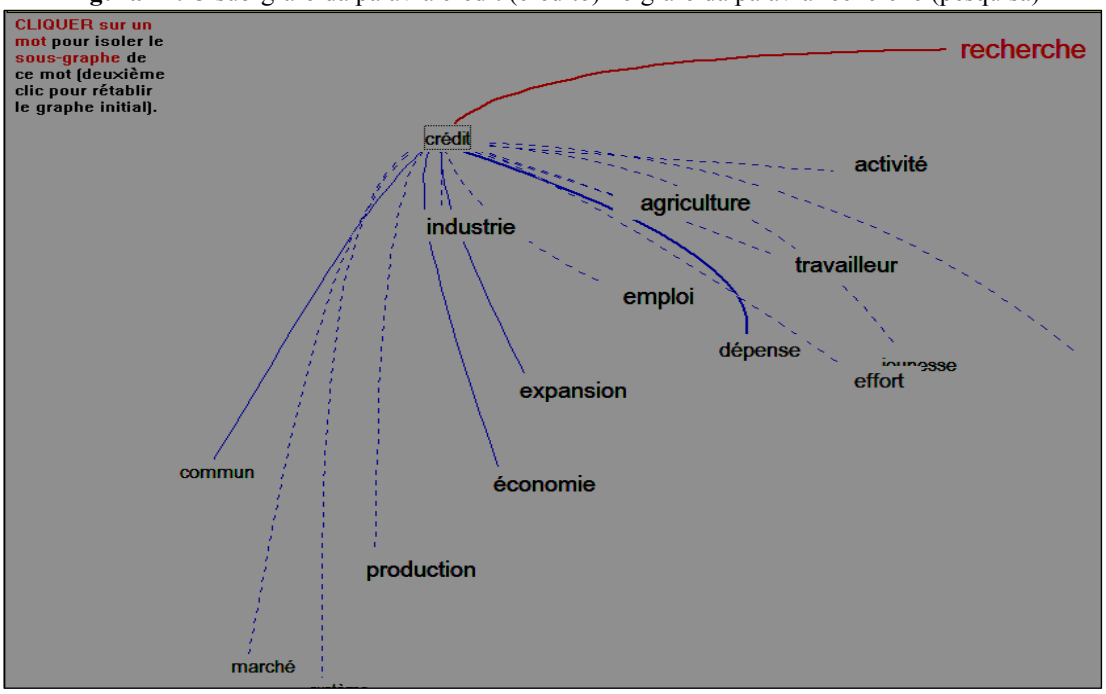


3 – Propomos aqui um recurso a mais para simplificar e esclarecer a interpretação dos grafos, que consiste em apresentar um **subgrafo** no qual somente estão representadas as relações que se vinculam a um único nó. Se isolarmos só o polo, somente os arcos vermelhos serão

[********] Nota do tradutor: sobre as análises em árvores, ver bibliografia de Xuan Luong – http://bcl.cnrs.fr/rubrique191.

conservados. Se se trata de um outro nó, a cor vermelha em geral desaparece e permanece só na relação com o polo. Os traços que levam ao nó designado por um clique não permanecem (como se vê com relação à palavra *crédit*, no subgrafo abaixo). Uma obrigação, no entanto, permanece: como na análise em árvores que precede, não devemos sair do círculo existente em torno do polo. Para escapar desta obrigação, basta elevar o nó selecionado à categoria de polo, mexendo assim com a hierarquia.

Figura 12. O sub-grafo da palavra crédit (crédito) no grafo da palavra recherche (pesquisa)



4 O PROGRAMA EXTERNO GEPHI*********

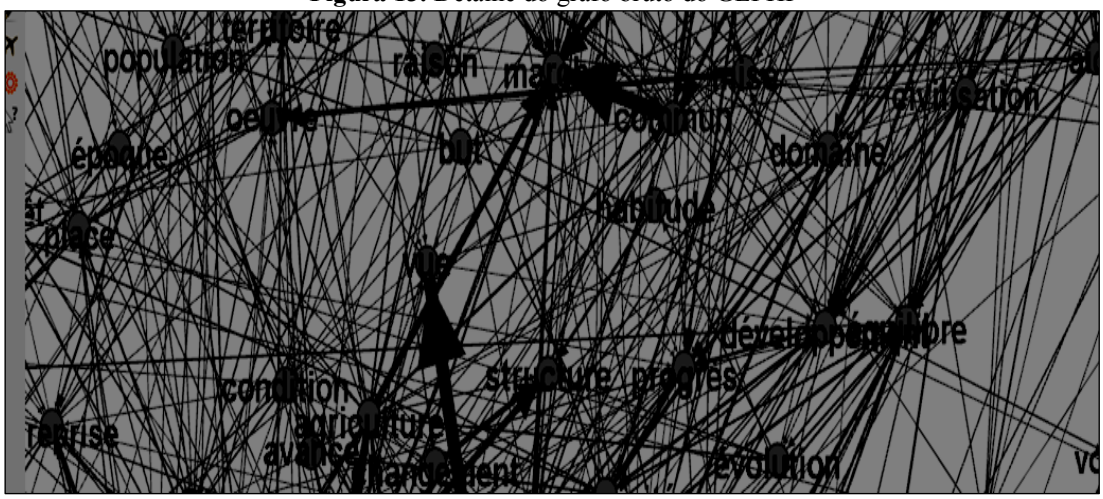
A função zoom tem dois movimentos opostos. Numa direção, ela serve para dar mais visibilidade ao essencial e para isolar os detalhes. Na outra direção, ela permite que se tenha uma visão da totalidade do objeto. Tentamos, sem dúvida, representar o mapa global da coocorrência, numa análise de fatores que encontramos no início deste texto. Mesmo com algumas poucas cores, a representação continua plana, sem relevo. Ora, há um programa aberto *(open source)* que oferece esta possibilidade de explorar as informações elementares fornecidas pelo programa Hyperbase com a indicação do nome e do número de cada nó e, para cada uma das relações *(edge)*, o número dos nós ligados entre si e a força desta relação (é

********* Nota do tradutor: Sobre o GEPHI, consulte, em português: <https://humanidadesdigitais.org/2013/08/16/analise-e-visualizacao-de-redes-o-gephi/>.

o desvio que se obtém com o cálculo hipergeométrico).

Quando solicitamos o botão GRAPHE TOTAL (ou GRAPHE GEPHI), estas informações são transmitidas num arquivo formato GML (o arquivo criado tem o nome da base associado ao sufixo .GML). Os dados são então comunicados ao programa externo GEPHI.EXE para que se obtenha o ***graphe* total** das coocorrências, e não mais o grafo de uma palavra somente. Temos, então, um grafo dinâmico, que se desenvolve de modo espetacular e de que representamos abaixo um detalhe somente.

Figura 13. Detalhe do grafo bruto do GEPHI

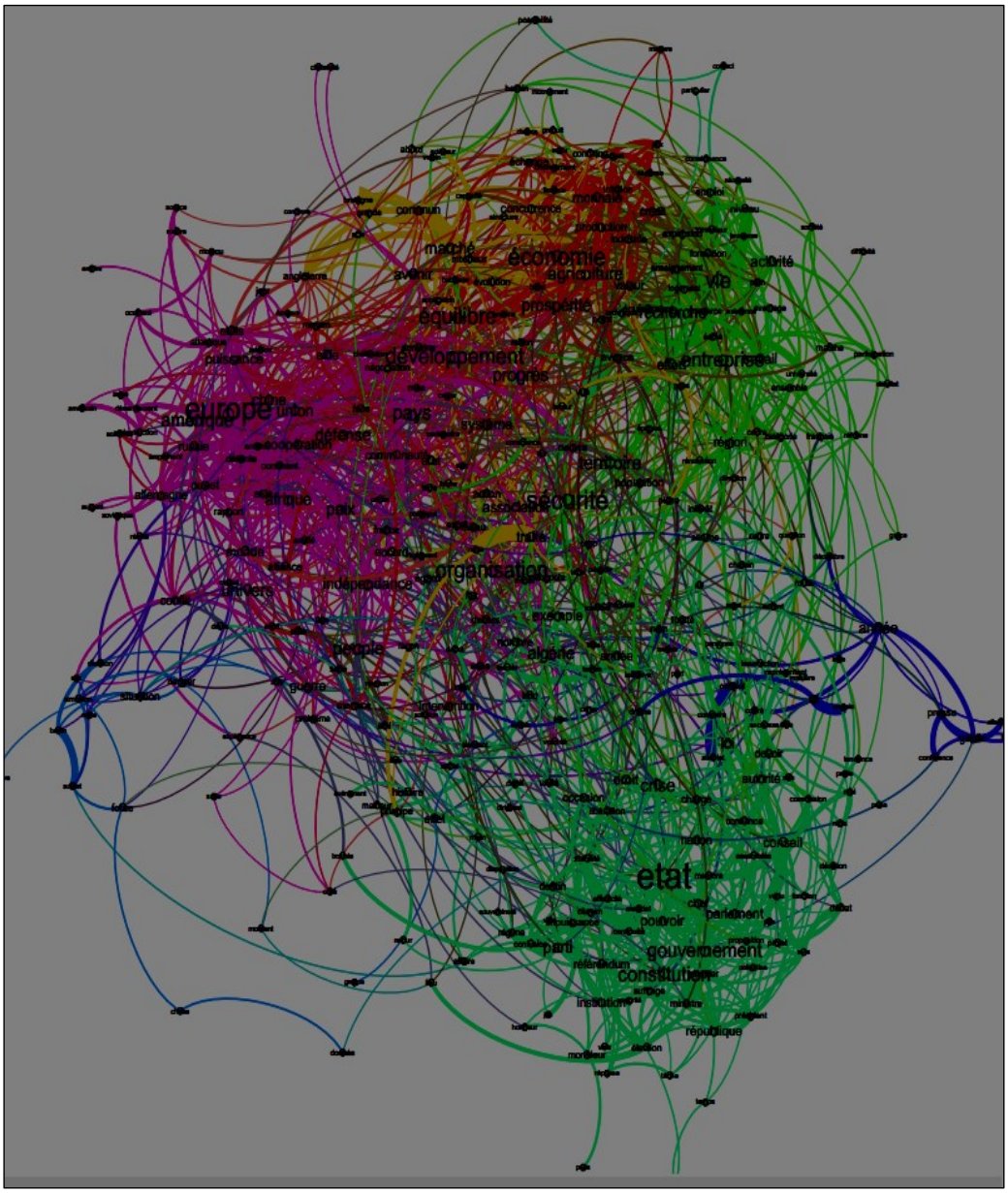


Vê-se quase que de imediato que uma espécie de poderosa atração põe a *agriculture* no espaço de atuação do *marché commun*. Mas é preciso ainda ajustar os numerosos parâmetros que determinam os cálculos e todas as modalidades que determinam também a sua execução. Se o grafo, nas suas grandes linhas, aparece bem depressa, é preciso ter paciência para, de modo preciso, dar a ele a clareza de que necessita. Várias etapas devem ser para isso obedecidas, nesta ordem:

- a **espacialização**, isto é, a força que movimenta os pontos dentro do espaço considerado; trata-se geralmente da força chamada *Atlas*, que trabalha com os diferentes ajustes de repulsão, de estabilização e de gravidade;
- os cálculos **estatísticos**, particularmente com a ativação das rubricas *Diamètre* (Diâmetro) e *Modularité* (Modularidade);
- a **classificação**, que determina os tipos de traços, a cor e o peso dos nós;
- e, finalmente, a **partição**, que trabalha com a repartição das cores e dá ao gráfico a sua lisibilidade.

É preciso, ainda, afinar o **aperçu** (**esboço**) que vai ser exportado para um arquivo PDF. O resultado é um grafo total em que todos os nós e todos os arcos estão representados; o algoritmo atribui a cada um deles um nome, um lugar, um traçado, uma espessura e uma cor. Apresentamos a seguir a figura obtida a partir dos discursos de De Gaulle. Neste caso também, as palavras aparecem numa espécie de triângulo: na parte inferior, onde domina a cor azul, está o léxico institucional; em cima, à esquerda, onde domina a cor vermelha, temos a política externa, enquanto que a economia ocupa o último terço, na parte superior, à direita.

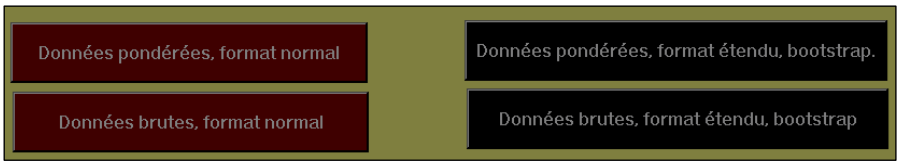
Figura 14. O grafo final de GEPHI relativo ao corpus dos discursos de De Gaulle



5 O PROGRAMA EXTERNO BOOTSTRAP DE L. LEBART

O programa Hyperbase utilizou durante muito tempo o programa de análise de fatores conhecido como TABET e gratuitamente cedido por um dos seus autores, André Salem[**********], que concebeu o muito conhecido programa LEXICO. Um programa de concepção similar, escrito também em linguagem Fortran, está por outro lado disponível para os casos em que o número de textos (isto é, de colunas) ultrapassa os limites da nossa versão do TABET, ou seja, 75. Concebido por Ludovic Lebart e integrado ao programa DTM-VIC, este programa contém uma **função de validação** (ou Bootstrap) que mede o grau de confiança que se pode razoavelmente dar aos resultados da análise. Para maiores informações, sugerimos a leitura das publicações de Lebart, particularmente do manual de instruções do seu programa DTM-VIC, que pode ser baixado livremente, com exemplos e com o próprio manual, no endereço http://www.dtmvic.com/05_SoftwareE.html[**********]. A escolha faz-se assim entre um método simples e um procedimento mais elaborado, dotado de complementos probatórios. No primeiro caso, nos vamos contentar com um formato reduzido; no segundo caso, vamos ocupar todo o ecrã. Acrescenta-se a isto uma opção relativa à preparação dos dados, com uma análise que leva em consideração as frequências brutas ou os dados ponderados. Neste último caso, os resultados são menos sensíveis às desigualdades existentes nas linhas ou nas colunas, e o gráfico, então mais arredondado, resiste melhor aos efeitos centrípetos ou centrífugos devidos às diferenças de extensão.

Figura 15. As opções da análise de fatores

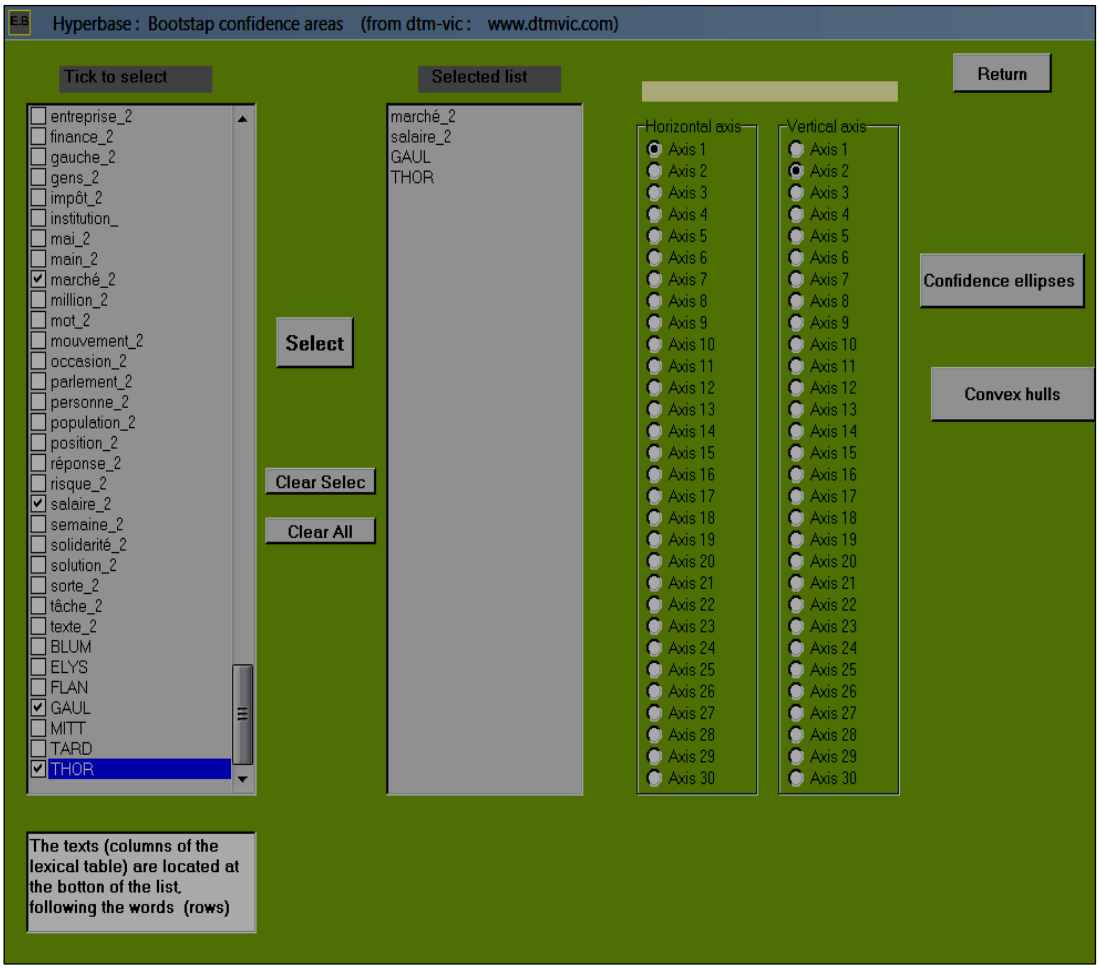


Ludovic Lebart deu-nos, com efeito, o seu programa de análise de fatores CORAN2.EXE e o programa complementar de testes probatórios ELLIPSE2.EXE e, além disso, arrumou, de acordo com o que era para nós mais conveniente, o diálogo com o usuário, tal como aparece na figura 16.

********** Nota do tradutor: Página pessoal de André Salem: <http://www.tal.univ-paris3.fr/asalem/>.

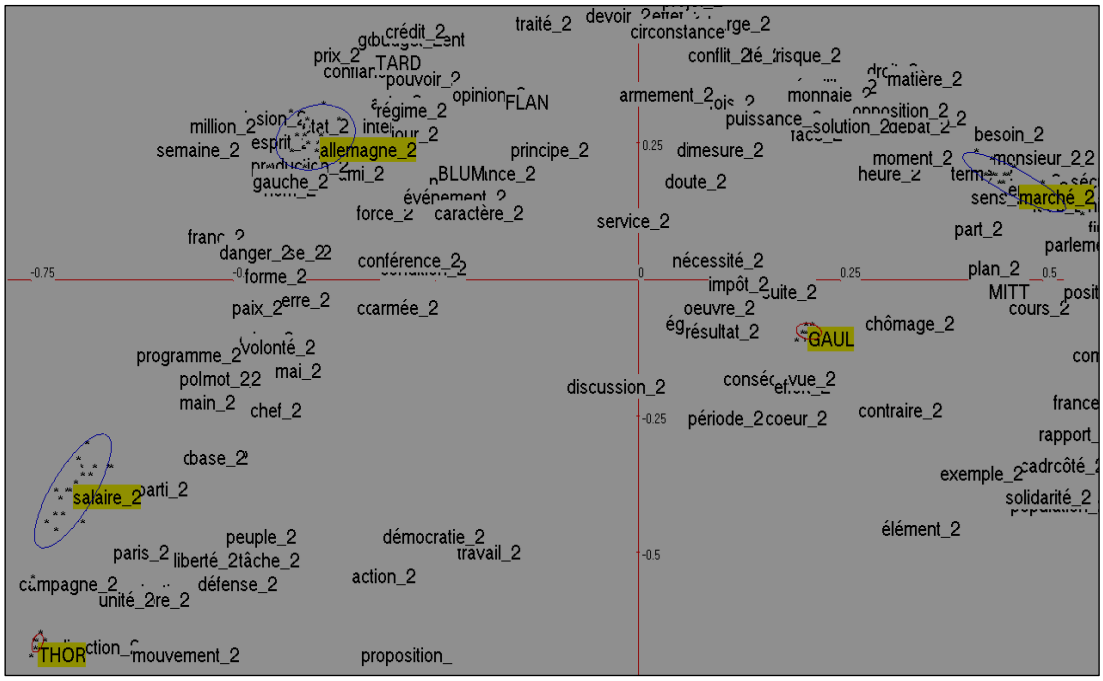
********** Nota do tradutor: Dados textuais – Visualização, Inferência, Classificação – os dados relativos a Ludovic Lebart podem ser consultados também em português – <http://www.dtmvic.com/01_mainP.htm>.

Figura 16. A utilização do programa de testes probatórios (Bootstrap partiel)



Como a análise de correspondências trabalha ao mesmo tempo com as linhas e com as colunas, cabe que sejam examinadas tanto as palavras (ou linhas) quanto os textos ou *corpora* (as colunas). Para chegar a estes últimos, é preciso primeiramente promover o desfile das palavras da lista, na parte esquerda da tela ou ecrã. É preciso assinalar, preenchendo as casas, de modo a fazer a seleção, escolhendo os elementos que devem ser submetidos aos testes probatórios. Os dois botões que se encontram à direita dão início a esta operação e a diferença está no tipo de apresentação que se dá aos resultados, que aparecem tanto em elipses quanto em polígonos.

Figura 17. Aplicação do Bootstrap partiel à análise de fatores



O programa pinta de amarelo os *corpora* e as palavras submetidos ao teste. A sua posição no gráfico não é mais um ponto único, mas, antes, uma zona mais ou menos concentrada ou que se dispersa, carregando consigo as marcas das quinze ou vinte localizações do ponto de que se trata se o cálculo for repetido um número igual de vezes. Se a análise não é muito segura e se o acaso se manifesta com mais força, a elipse que contém estes pontos incha até transbordar, atingindo diferentes quadrantes do gráfico, como se se tratasse de uma arma mal ajustada que vai, assim, atingir o alvo em vários pontos. Vê-se que não se trata disso no exemplo aqui proposto. Os dois *corpora*, THOREZ à esquerda e GAULLE à direita, foram ambos atingidos com tanta precisão que todos os impactos estão reunidos em alguns poucos milímetros. A margem de erro é só um pouco mais estreita para as palavras analisadas, *salaire* (salário), *Allemagne* (Alemanha) e *marché*. A elipse que as envolve, mesmo alongada, nem mesmo tem 2 centímetros, o que é pouca coisa num gráfico que ocupa 50 centímetros no ecrã.

Acrescentemos a isto que os resultados, para a mesma análise, têm o mesmo valor probatório quando a escolha diz respeito a outras palavras ou a outros *corpora*. Mas não devemos a partir daí concluir que o teste é inútil. Tudo depende da quantidade de dados disponíveis. Quando os dados compreendem, como é aqui o caso, milhões de observações, a parte que caberia ao acaso encolhe e fica bem mais modesta, e é o que pode ser constatado na presente análise, onde podemos ver que os resultados são mais precisos para os *corpora* do que para as

palavras, simplesmente porque os *corpora* têm uma extensão bem mais importante.

6 OS PROGRAMAS EXTERNOS ALCESTE E IRAMUTEQ

Há já dez anos que o programa Hyperbase propõe uma relação com o programa ALCESTE, ao qual ele cede os seus dados para receber em troca também ensinamentos. ALCESTE baseia-se num algoritmo original e muito eficaz, que Max Reinert********** criou e ajustou há cerca de vinte anos. Mas uma nova montagem deste algoritmo, realizada por Pierre Ratinaud com o nome de IRAMUTEQ, surgiu ultimamente, numa plataforma diferente, que aceita dados com o mesmo formato. Novas possibilidades de processamento foram, assim, abertas para os nossos dados, com os dois programas, já que o arquivo transmitido foi o mesmo.

Este arquivo é, no entanto, diferente daqueles que estes programas normalmente processam ou analisam. A sua matéria-prima é constituída de textos em formato ASCII, sem balizas próprias, a não ser aquelas "com estrelas", que designam o início de um texto novo e descrevem sumariamente os seus parâmetros ou palavras-chave. Antes de aplicar a estatística a estes dados, devem ser aplicados pesados processamentos documentais e linguísticos, análogos aos que o programa HYPERBASE também, por outro lado, faz: indexação, lematização, consulta dos dicionários, processamento das homografias... Consideramos, então, que era melhor evitar todo este trabalho, já que o programa Hyperbase se tinha antes encarregado disso. Os dados que são transmitidos são então dados de textos já desambiguizados, lematizados e já sem uma qualquer marca flexional. Como o que se espera obter é de ordem temática e semântica, deixamos de lado palavras gramaticais, e até mesmo advérbios, verbos e adjetivos (salvo no caso de haver uma prévia intenção de trabalhar com estas categorias). O texto transmitido é então um texto **radicalmente limpo:** ele só contém os 300 ou 400 substantivos cuja lista foi estabelecida quando do processamento das coocorrências; foi no exato momento deste processamento que o arquivo destinado aos programas Alceste e Iramuteq foi criado. Eis as primeiras linhas do arquivo GAULL1.TXT (cujo nome é o nome da base, sendo que o 1 substitui a última letra):

*****1958*
allocution année gaulle juin unité
(alocução ano gaulle junho unidade)

********** Nota do tradutor: sobre Max Reinert e Alceste: https://fr.wikipedia.org/wiki/Alceste_(logiciel).

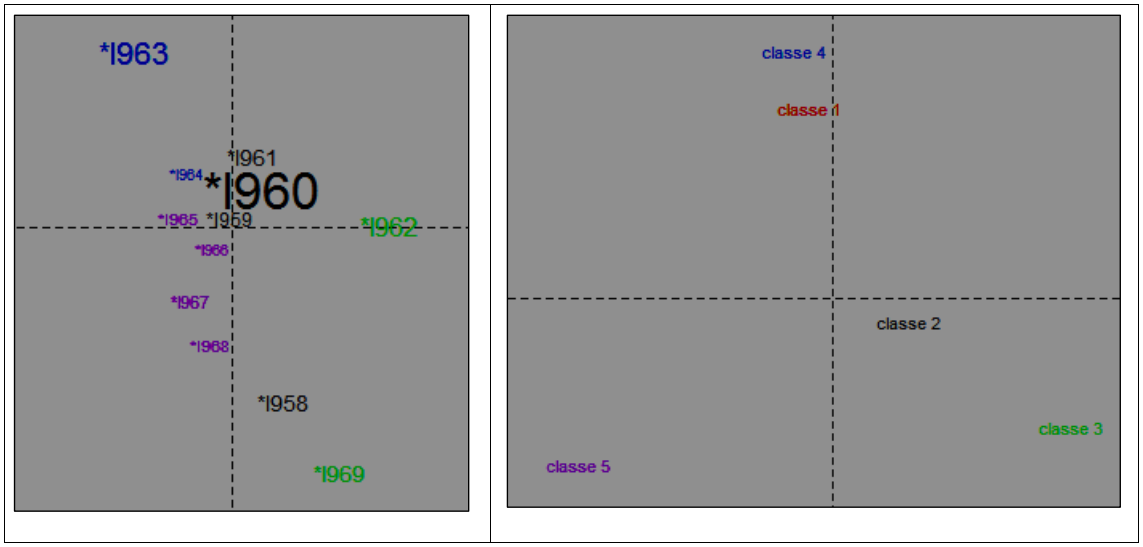
france monde oeil point
(França mundo olho ponto)
charge pays
(encargos país)
algérie armée avenir changement crise drame épreuve place
(argélia exército futuro mudança crise drama prova-provação lugar)
population
(população)
affaire an état milieu peuple régime univers
(negócio-caso ano estado meio povo regime universo)
gens pouvoir république valeur
(gente-pessoas poder república valor)
assemblée face fonction groupe mois parti problème
(assembleia face-diante função grupo mês partido problema)
afrique algérie coopération corps états france lien
(África argélia cooperação corpo estados França laço-vínculo)
métropole peuple
(metrópole povo)
rapport sorte
(relação sorte-destino)
action fois monde pays place sécurité vue
(ação vez mundo país lugar segurança vista)
économie effort équilibre faute finance grâce intérieure pays
(economia esforço equilíbrio falta-erro finanças graça interna país)

A primeira linha corresponde à primeira variável "com estrelas", com um único parâmetro: o nome do texto (neste caso o ano de 1958). As linhas seguintes correspondem às frases do discurso, sem todas aquelas palavras que foram expurgadas e que estão fora da lista. Como os intermediários foram eliminados, a coocorrência adquire aqui um caráter brutal e espetacular, de que se espera mais simplicidade, mais rapidez e mais precisão nos cálculos. E é isto que com efeito se pode observar: o processamento destes dados fica muito mais fácil, tanto com ALCESTE como com IRAMUTEQ. Uma precaução deve no entanto ser tomada para definir o tamanho dos segmentos (ou *uce*, unidades de contexto elementares): como a concentração é muito forte, depois que os acessórios foram eliminados, uma sequência de 7 a 10 palavras

corresponde a uma sequência de 40 palavras no texto original. Ao lançar o programa é preciso então fixar em 10 pelo menos o tamanho de cada *uce*, ou pelo menos respeitar as marcas de fim de frase.

Como os dois programas são muito completos, daremos somente uma ilustração: aquela que tem por base o algoritmo de Max Reinert e que divide os textos e as palavras segundo as classes a que pertencem e de que o programa determinou os contornos e o conteúdo. Estas classes têm um espaço, uma cor e um número. Mas o seu nome, como o nome dos "fatores" na análise de fatores, fica reservado para a interpretação. No exemplo de De Gaulle (figura 18), temos as duas representações: a dos textos (que é a dos anos), à esquerda, e a das classes, à direita, que se sobrepõem e cujas cores correspondem. A cor magenta, da classe 5, aparece no mesmo quadrante que os anos 1965, 1966, 1967 e 1968. No centro, os primeiros anos, que vão de 1958 a 1961, são da classe 2. A classe 4 diz respeito aos anos de 1963 e 1964, que aparecem no quadrante superior esquerdo, enquanto que os textos restantes aparecem no quadrante oposto ao da classe 2. É visivelmente o tempo que determina a sucessão das classes, já que cada uma delas (ou quase) engloba uma etapa cronológica homogênea.

Figura 18. A análise dos textos (ou dos anos)e das classes

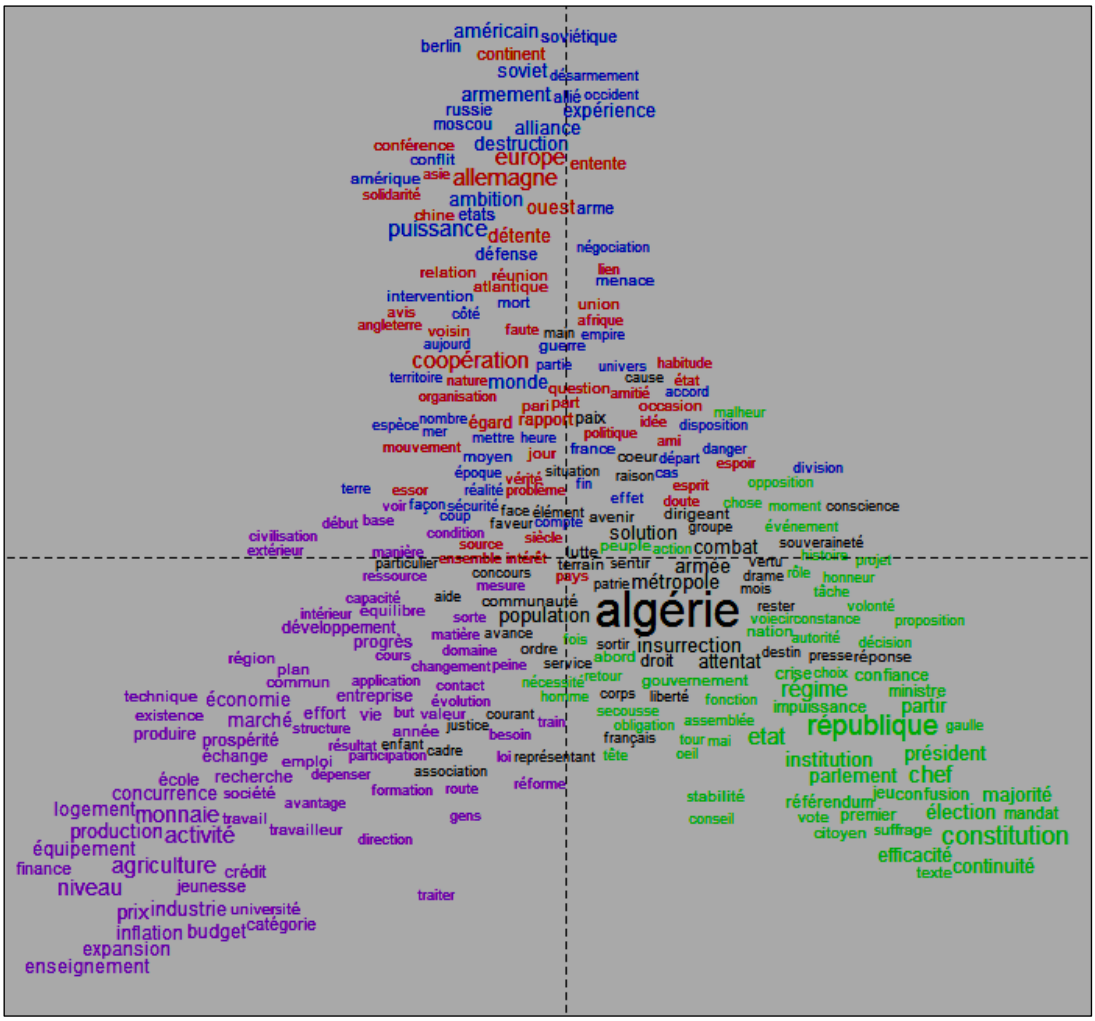


Mas é a representação das palavras (figura 19) que nos manda a mensagem mais clara. Reconhecemos ali a forma triangular já encontrada em análises anteriores, com três pontas: a política externa está na parte superior, a gestão econômica está à esquerda e as questões institucionais estão à direita. O futuro da Argélia se resolve no centro, misturado com as questões de ordem interna. A política externa aparece, por outro lado, em duas classes

distintas: uma, de cor vermelha, diz respeito à Europa; a outra, de cor azul, diz respeito ao resto do mundo.

E podemos aceitar esta apresentação muito clara dos temas e dos discursos de De Gaulle. Mas o espectro das cores poderia ser analisado com nuanças mais finas como também mais grosseiras. Isto depende de certos parâmetros que o utente escolhe como quiser. Se ele deixar de fazer isso, as opções pré-definidas decidem sem que ele participe. No caso em pauta, deixamos que o programa IRAMUTEQ fizesse a escolha. Mas se propusermos ao programa ALCESTE os mesmos dados, as opções pré-definidas produzem uma análise um pouco diferente, com sete classes em vez de cinco. A política externa faz das questões de desarmamento uma classe completamente separada. E, em termos de política interna, uma separação se faz entre as reformas e as instituições.

Figura 19. A análise das palavras e dos temas nos discursos de De Gaulle
(segundo o programa IRAMUTEQ)



7 CONCLUSÃO

A idade do autor não permite desde já esperar grandes evoluções. O envelhecimento da linguagem utilizada (Toolbook versão 10.5) também segue pelos mesmos caminhos. Devemos, então, considerar a versão 10 como o último avatar de um produto que foi talvez útil para algumas pessoas. Convidamos então agora todos aqueles que se sentem eventualmente decepcionados com os limites atingidos a olhar para as realizações mais recentes, como os programas TXM e IRAMUTEQ e, sobretudo, a acompanhar os progressos de uma **versão Internet** do programa HYPERBASE, realizada por Laurent Vanni no sítio do laboratório (logometrie.unice.fr).

Recebido: 10 de julho de 2017
Aceito: 12 de julho de 2017