Analyse thématique du discours : segmentation, structuration, description et représentation

Nicolas Hernandez, Brigitte Grau

▶ **To cite this version:**

Nicolas Hernandez, Brigitte Grau. Analyse thématique du discours : segmentation, structuration, description et représentation. Conférence CIDE'05, 2002, Hammamet, Tunisie. hal-02456869

HAL Id: hal-02456869
https://hal.science/hal-02456869

Submitted on 27 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse thématique du discours : segmentation, structuration, description et représentation

Nicolas Hernandez, Brigitte Grau

LIMSI/CNRS, BP 13,
F-91403 ORSAY (France)
[Hernandz|Grau]@limsi.fr

RESUME :

L'analyse thématique est une étape importante pour de nombreuses activités en traitement automatique des langues, telles que le résumé ou la recherche d'information par exemple. Dans ce papier, nous présentons trois phases d'analyse : la segmentation thématique, la mise en évidence d'une structure de texte, et la description des segments par identification de leur thème et de leur rôle. Dans un objectif applicatif de navigation intra-document, nous présentons aussi un schème d'annotation générique multi-annotation fondé sur la technologie XLink qui nous sert de support de description des données.

MOTS CLEFS : segmentation thématique, structuration du discours, identification des thèmes, rôle des segments, schème d'annotation XML, navigation, visualisation, XLink.

ABSTRACT:

Thematic analysis of discourse can be extremely useful to natural language processing applications such as automatic text summarization or information retrieval. In this paper, we propose a three-step analysis: topic segmentation, topic structure analysis and finally segment description by topic identification and role classification. Led by a browsing application, we propose an XML meta-scheme for multi-annotations based on XLink technology to describe our data.

KEYWORDS : Thematic analysis, text segmentation, discourse structure, topic identification, text classification, annotation scheme, XML, browsing, visualization, XLink.

1. Introduction

L'analyse thématique est une étape importante pour de nombreuses activités en traitement automatique des langues, telles que le résumé ou la recherche d'information par exemple (Hovy, 2001).

En ce qui nous concerne, nous situons notre travail dans le cadre applicatif de la visualisation de documents textuels et de la navigation intra-document (Boguraev and al., 1999). Bien que conduits par ces objectifs, nos développements se trouvent principalement en amont de toutes chaînes de traitement. En effet, l'un des problèmes que nous relevons dans la présentation de documents électroniques via Internet est que la plupart de ces documents obéissent à une mise en page de documents papiers. Cette remarque est à l'initiative de notre projet et nous amène à proposer différentes analyses pour permettre un reformatage dynamique des documents (voir le schéma d'ensemble). Outre les problématiques de structuration du discours, nous allons plus loin en proposant un mode de présentation de l'information pour une visualisation plus ciblée des documents. Notre corpus d'étude est composé de textes de type scientifique et expositif.

Les sections suivantes présentent en détail les différentes phases de notre analyse. Dans la section 2, nous présentons notre méthode de segmentation thématique fondée sur la répartition des mots. Celle-ci nous sert de base pour notre méthode de structuration pour une organisation des textes « en boîtes » - section 3. La section 4 recense nos techniques de description de segments. Et enfin, la section 5, comme un support à toutes ces parties, propose un schème d'annotation générique fédérateur à tous les types de données que nous manipulons et à visée applicative, la navigation intra-document.

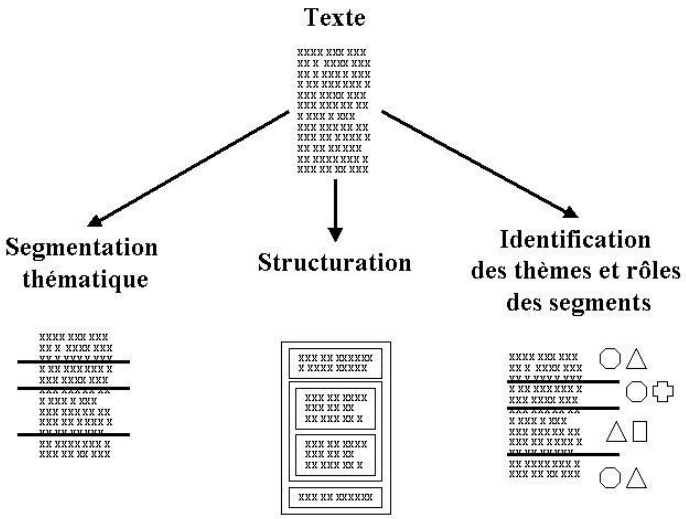


Schéma d'ensemble

2. Segmentation thématique

Une analyse thématique fondée sur la répartition des mots dans le texte part du constat que le développement d'un thème entraîne la reprise de termes spécifiques, notamment lorsqu'il s'agit de textes techniques ou scientifiques. La reconnaissance de parties de texte liées à un même sujet est alors fondée sur la distribution des mots et leur récurrence. Si un mot apparaît souvent dans l'ensemble du texte, il est peu significatif, alors que sa répétition dans une zone limitée est très significative pour caractériser le thème du segment. Le principe général appliqué par les différents systèmes reposant sur ce critère de délimitation de thème (Masson, 1995 ; Salton, 1996 ; Hearst, 1997) consiste à associer un vecteur de descripteurs à une zone de texte, où les descripteurs sont les mots pleins lemmatisés du texte (nom, adjectif et verbe) et leurs valeurs le nombre d'occurrences, pondéré, dans la zone. La pondération que nous utilisons est de type *tf.idf*, où la fréquence d'un mot est divisée par le nombre de paragraphes qui le contiennent. Le produit scalaire de ces vecteurs permet ensuite de regrouper ou de séparer les zones qu'ils décrivent selon qu'ils sont proches ou non.

Nous avons choisi le paragraphe comme unité de base, comme dans Masson, alors que Hearst délimite des blocs de pseudo-phrases, composées d'un nombre fixé de mots. Le choix du paragraphe comme unité minimale se justifie par le fait que les auteurs exposent en général un point de vue par paragraphe et ont tendance à introduire des paragraphes pour aérer le texte plutôt qu'à regrouper plusieurs sujets en un bloc. Pour déterminer si des zones de texte contiguës sont liées ou non, nous construisons une courbe dont les valeurs indiquent des ruptures thématiques lorsqu'elles sont situés sous un seuil, ce seuil étant fixé en fonction des valeurs de la courbe. Après expérimentations (Masson, 1998), il a été fixé à la moyenne des valeurs moins la moitié de l'écart type.

Ce type d'analyse donne de bons résultats sur des textes dont les termes caractéristiques des thèmes développés ne possèdent pas de synonymes et sont donc repris dans les segments. Mais en ce qui concerne les textes du type article de journaux, ce type de méthode tend à ne jamais trouver de liaison entre unités de base de l'analyse (Ferret, 1998).

Aussi d'autres méthodes reposant sur la cohésion lexicale et utilisant une source de connaissances non dédiée peuvent être employées, permettant alors de délimiter des segments thématiques sans fixer de taille minimale, par exemple (Ferret, 1998 ; Kozima, 1993).

Quelle que soit la méthode utilisée, la segmentation pose des marques de début et fin de segment dans le texte, une fin de segment étant toujours suivie par une marque de début, sauf pour le dernier segment.

3. Structuration

Lorsqu'un auteur traite un sujet, il en expose en général un point de vue, en développe des aspects particuliers, ce qui conduit à délimiter des segments distincts, et peut ensuite revenir au propos initial. Cependant les enchaînements possibles de

différentes thématiques suivent des critères forts de cohérence. Par exemple, lorsqu'on effectue une digression, puis que l'on retourne au sujet qui la précédait, il est rare de retrouver le thème de la digression par la suite. Le texte suit ainsi une organisation que l'on appellera « emboîtée ». Ce type de structure est très courant, surtout dans les textes expositifs, même si on peut trouver des structures dites « entrelacées », où plusieurs thèmes sont développés en parallèle. Ce dernier style se trouve plutôt dans des articles de journaux.

Peu de travaux portent sur l'explicitation automatique de la structure d'un texte. Certains requièrent une analyse conceptuelle complète des phrases du texte, comme par exemple (Hobbs and al., 1993) ou (Asher and Lascarides, 1994) qui explicitent des relations causales et temporelles entre phrases, ce qui est irréalisable pour traiter des textes non restreints. D'autres ne proposent pas une théorie suffisamment bien spécifiée pour envisager sa mise en œuvre ; il en est ainsi des modèles de (Grosz and Sidner, 1986) quant à l'explicitation des intentions dans le discours et (Mann and Thompson, 1988) avec la RST. (Marcu, 2000) propose la construction automatique de la structure d'un texte, fondée sur la RST. Les relations sont explicitées par l'utilisation d'indices de surface, aussi bien des indices linguistiques que des indices de cohésion lexicale. Cependant, quelles que soient les marques utilisées, Marcu met en relation des segments consécutifs et construit une structure arborescente, comme le fait (Yaari, 1997) en rapprochant aussi les segments consécutifs sur un critère de similarité lexicale. (Salton and al., 1996) utilise aussi une mesure de similarité lexicale en construisant un graphe des relations entre segments afin de mettre en évidence les thèmes d'un texte, définis comme certains chemins caractéristiques dans le graphe, pouvant relier des segments discontinus.

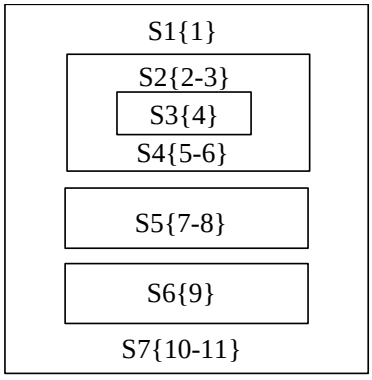
Afin de construire la structure emboîtée d'un texte, nous cherchons à mettre en évidence d'abord le niveau le plus englobant en recherchant les deux segments non consécutifs les plus liés. On délimite ainsi la portée du thème développé dans ces passages. Nous ré-appliquons récursivement le même principe aux segments inclus dans la structure de plus haut niveau ainsi qu'à ceux qui sont restés au même niveau.

La mesure de liaison de deux segments est la même que celle que nous appliquons pour segmenter le texte et qui est appliquée sur les unités de base. Un segment étant constitué du regroupement de plusieurs unités consécutives, ses descripteurs sont l'union des descripteurs de ces unités (i.e. l'ensemble de tous les mots retenus), leur poids étant la moyenne des poids de chaque unité de base. Nous avons choisi la moyenne des valeurs afin de conserver des éléments comparables et ne pas déséquilibrer la description d'un thème, qu'il soit développé dans une seule unité de base ou dans un regroupement de plusieurs unités. La décision de lier ou non deux segments non consécutifs suit le même principe que pour la segmentation, aussi le seuil de coupure est le même.

Prenons un exemple. La segmentation du texte « Le vin jaune », provenant du magazine « Pour la science, octobre 1994 » (cf. annexe) et possédant 11 paragraphes, conduit à la délimitation des 7 segments suivants, où un segment Si est formé de la concaténation des paragraphes j à k entre accolades :

S1{1}, S2{2-3}, S3{4}, S4{5-6}, S5{7-8}, S6{9}, S7{10-11}

La structure qui en est extraite est la suivante :



La liaison trouvée entre le premier est le dernier segment est classique et relie l'introduction et la conclusion. On peut aussi constater que parmi les segments emboîtés, un niveau supplémentaire a été détecté, avec la liaison des segments 2 et 4, entraînant donc l'imbrication de S3. Nous avons appliqué cette structuration sur différents textes scientifiques, avec des résultats satisfaisants. Nous travaillons actuellement à l'élaboration d'un protocole d'évaluation systématique, un tel cadre n'existant pas actuellement.

4. Description des segments textuels

Plusieurs axes de recherche s'intéressent aux moyens de décrire un texte ou une de ses portions, certains par clustering et classification (Ferret and Grau, 1998), d'autres en extrayant les phrases les plus significatives (Mani and Maybury, 1999), d'autres encore en identifiant les thèmes significatifs.

La majorité des auteurs travaillant sur l'identification de thèmes s'appuie sur les groupes nominaux pour décrire des thèmes d'un texte (Mather and Note, 2000 ; Boguraev and Kennedy, 1997 ; Lin and Hovy, 1997) se différenciant principalement dans leur manière de mesurer la proéminence thématique de ces expressions. Ainsi pour certains, la pertinence de ces expressions dépend de leurs mises en valeur par des marques linguistiques (Porhiel, 2001), pour d'autres en connaissance du genre du document, de leurs positions relatives et absolues (Lin and Hovy, 1997), et pour d'autres enfin, plus classique, de leurs fréquence et de leurs distributions (Boguraev and Kennedy, 1997).

Les travaux de (Boguraev and Kennedy, 1997 ; Boguraev and al., 1999) sont les plus avancés dans le domaine. Ils proposent ainsi de présenter les segments de texte par les groupes nominaux représentatifs des entités les plus significatives du segment considéré, accompagnés de spécifications contextuelles de différents niveaux de granularité (groupe verbal, proposition minimale, phrase).

Notre travail s'inspire de ces travaux de part un objectif commun : obtenir une description intelligible plus fine que la simple présentation de phrases extraites tout en n'introduisant pas de procédés de génération automatique. Pour atteindre ce

niveau de finesse, nous regardons les textes au niveau des entités qu'ils mettent en contexte. Nous distinguons différents types d'entités : celles à propriété thématique et celles à propriété générique (les premières étant encore subdivisées en globales et locales). Par rapport à notre corpus de nature scientifique, les entités thématiques sont propres à un document, elles correspondent abusivement à « ce dont parle l'auteur » ; le lexique générique quant à lui se situe sur un autre axe et correspond au vocabulaire propre à l'ensemble des documents de nature scientifique (mis à part les mots vides), c'est-à-dire par exemple « hypothèse, analyse, problématique, … ».

L'importance thématique d'une entité est calculée en fonction de sa fréquence et de sa distribution (mesure *tf.idf*), et ce suivant différents niveaux d'observation : local, par rapport à chaque segment thématique, et global, par rapport au document entier. Les groupes nominaux sont identifiés comme désignant une même entité par rapprochement lexical. Afin de rendre les valeurs de fréquence et de distribution plus précises, nous considérons comme antécédents des anaphores pronominales (« il, elle, ils et elles ») le premier groupe nominal qui précède et vérifie le genre et le nombre du pronom.

Bien que la reconnaissance des entités génériques soit à réaliser pour chaque document, leur caractérisation dépend d'un seul pré-traitement de notre corpus. En effet, cette base de mots génériques est construite en prenant l'intersection des mots pleins (noms, verbes et adjectifs) de tous les documents de notre corpus. Chaque mot plein est réduit à une seule occurrence par document et seulement les mots dont la fréquence est supérieure à la moyenne du nombre de documents sont conservés.

Ainsi, la visualisation de texte pourra suivre les différents degrés de développement thématique dégagés lors de la structuration thématique et la description d'un segment reposera sur la présentation de ces différentes entités et sur leur catégorisation (pertinence globale, pertinence locale, lexique générique). Dans l'état de nos avancées, nous émettons l'hypothèse que les étiquettes génériques suffisent à décrire le rôle du segment où elles apparaissent. Nous envisageons de complexifier plus tard nos motifs de recherche en s'inspirant des travaux réalisés dans le domaine de l'extraction d'information notamment par la méthode d'exploration contextuelle (Ben Hazez and al., 2001).

5. Représentation XML des données

Fondamentalement toutes les informations que nous calculons à partir du texte sont destinées à produire des annotations pour enrichir le texte. Initialement, nous voulions définir un schème d'annotation pour faciliter des échanges de résultats entre différents modules de notre chaîne de traitement. Nous nous sommes rendus compte de l'importance de définir un schéma d'annotation cohérent et ce principalement pour les retombées applicatives, que cela soit pour justifier nos développements ou pour des objectifs d'applications finales de navigation et de visualisation.

Notre cahier des charges reprenait les contraintes suivantes :

- Une forte expressivité afin de représenter tous les objets manipulés aussi bien les structures de base (token, phrase, paragraphe) que

plus complexes (thèmes, les segments thématiques, la structure thématique). Nous devions aussi pouvoir exprimer les relations que ces différents objets entretiennent entre eux, et finalement décrire à la fois ces objets et ces relations avec des référents potentiellement différents.

- Une optique fédératrice entre les différentes équipes de recherche intervenant dans la réalisation de notre chaîne de traitement.
- Un objectif final d'application à des tâches de visualisation de textes et de navigation intra-texte.

Notre schème de représentation se veut générique, multi-annotation, et autorisant des structures en graphe. De nombreux auteurs ont cherché à apporter des réponses à ces différents objectifs.

Ainsi (Cristea and al., 1998) souligne l'intérêt d'appliquer un principe de séparation entre les éléments relationnels et les éléments objets, et ce dans le cadre applicatif de la représentation de la structure rhétorique des textes et de la coréférence. Ce principe est fondamental pour représenter les structures en graphe. Pour représenter les liens ils s'inspirent des directives du TEI (Sperberg-McQueen and Burnard, 1994). Cependant leur schème est spécifique à la tâche qu'ils traitent.

Le schème XCES (Ide and al., 2000) est une implémentation XML du CES (Corpus Encoding Standard), celui-ci fait partie des directives d'EAGLES et concerne l'annotation linguistique de données textuelles. XCES n'est pas encore figé, mais dans un exemple de sa forme potentielle il apparaît comme un schème générique avec de multiples annotations et des références aux textes analysés par XPointer. Les relations annotées sont alors exprimées en combinant relations de contention père/fils des éléments XML et des couples d'attributs type ID/IDREFS. Défini comme tel, ce schème nous semble peu homogène de par ses représentations complexes de relation et limité en expressivité notamment pour les structures en graphe. Il présente néanmoins l'intérêt d'une réflexion sur la forme générique d'une annotation.

Les réflexions sur des cadres représentatifs multi-annotations (Lopez and Romary, 2000 ; Dybkjær and Bernsen, 2000) se rapprochent de ce que nous désirons tout en étant conformes aux directives du TEI. Néanmoins l'apparition de nouvelles technologies XML comme XLink (De Rose and al., 2001) et des outils qui commencent à apparaître pour les manipuler nous a fait adopter un autre schème d'annotation reposant sur la technologie XLink et XPointer.

Comme le souligne (Ide and al., 2000) le choix du schème d'annotation n'est pas dans l'absolu très important, dans la mesure où l'information est toujours présente. D'ailleurs, moyennant des lignes d'écriture, la transformation d'un document XML en un autre est chose possible grâce à XSL (XML Stylesheet Language). En définissant notre propre schème d'annotation, nous voulions éviter de forcer nos données à s'emboîter dans un schème pas nécessairement adapté. Cela peut à notre tour nous être reproché, mais notre principal intérêt dans la définition d'un schème générique réside dans l'utilisation d'une même API (Application Program Interface) lors de nos développements.

Finalement nous adoptons une annotation tripartite dissociant le texte annoté, la description des objets du texte que l'on annote et les relations entre les objets et le texte et entre les objets.

Le texte annoté contient les balises des structures de base (token, phrase, et paragraphe), repérées préalablement à partir d'indices typographiques. Pour être indépendant de tout balisage, l'idéal aurait été de se référer aux caractères pour repérer les objets du texte mais nos traitements ne nécessitent pas un tel niveau de granularité. Ou bien nous aurions pu exploiter la caractéristique de XPointer à désigner des token-mots sans faire intervenir des ancres (style HTML) pour les repérer. Néanmoins d'une manière générale il nous a semblé plus simple de considérer un balisage minimum pour faciliter certains traitements. Chacune des balises s'accompagne d'un identifiant unique.

```
<!-- exemple d'encodage d'un segment  -- >
<Struct id="seg3" type="thematicSegment">
   <Feature>
      <Name>EntiteThemeGlobal</Name>
      <Value>vin</Value>
   </Feature>
   <Feature>
      <Name>EntiteGenerique</Name>
      <Value>analyse</Value>
   </Feature>
</Struct>
```

```
<!- - exemple d'encodage d'un thème -- >
<Struct id="th1" type="theme">
   <Feature>
      <Name>LinguisticExpression</Name>
      <Value>Vin jaune</Value>
   </Feature>
   <Feature referencePoint="seg1">
      <Name>Frequence</Name>
      <Value>0.7</Value>
   </Feature>
</Struct>
```

Concernant la description des objets, l'annotation donne un identifiant, précise un type et, s'inspirant de (Ide and al., 2000 ; Grishman, 1998), fournit le moyen de représenter des couples : nom d'une caractéristique et valeur. Notre originalité vient de pouvoir spécifier (s'il y a lieu) dans quel référentiel une valeur existe. Cela s'avère utile par exemple lorsque l'on précise pour un thème donné son poids d'un point de vue global ou local.

```
< !-- Exemple d'une relation  décrivant un segment, seg4, traitant des thèmes th2 et th4 -->
<myLinks xmlns:xlink=http://www.w3.org/1999/xlink xmlns="http://my.schemeDAnnotation/syntax/">
      <myLinkingElements xlink:type="extended">
         <myResource xlink:type="locator"
                  xlink:href="segments.xml#seg4"
                  xlink:role="http://my.schemeDAnnotation/roles/segments"
                  xlink:label="seg4"/>
         <myResource xlink:type="locator"
                  xlink:href="themes.xml#th2"
                  xlink:role="http://my.schemeDAnnotation/roles/themes"
                  xlink:label="th2"/>
         <myResource xlink:type="locator"
                  xlink:href="themes.xml#th4"
                  xlink:role="http://my.schemeDAnnotation/roles/themes"
                  xlink:label="th4"/>

         <myLink xlink:type="arc"
                  xlink:arcrole="http://my.schemeDAnnotation/roles/isAbout"
                  xlink:from="seg4"     xlink:to="th2"/>
         <myLink xlink:type="arc"
                  xlink:arcrole="http://my.schemeDAnnotation/roles/isAbout"
                  xlink:from="seg4"     xlink:to="th4"/>
      </myLinkingElements>
</myLinks>
```

La description de nos liens repose sur le standard Xlink. Les liens sont considérés comme des éléments à part entière que l'on peut décrire sémantiquement à l'aide d'URI (Universal Resource Identifier). Outre ses qualités pour exprimer du multi-annotation muti-document (intra- et extra-) multi-sources et destinations, XLink présente aussi la caractéristique de décrire les comportements des liens HTML (quand un lien sait s'activer, ce qui se passe quand le lien est activé). Bien que cette technologie ne soit pas encore implémentée ni par Netscape, ni par Internet Explorer cette caractéristique est un atout majeur dans un objectif de navigation intra-texte.

6. Conclusion

En pratique, les différentes phases d'analyse énumérées ci-dessus donnent lieu à une architecture dont le prototype est en cours de développement. L'évaluation d'un tel système est très difficile. Nous l'envisageons de manière indirecte par la participation au projet REGAL (Résumés Guidés par les Attentes du Lecteur), action cognitique du CNRS mettant en collaboration les équipes LaLIC du CAMS, UMR LATTICE, CEA et LIR du LIMSI.

Les perspectives d'évolution de notre système sont nombreuses, notamment grâce à la souplesse de notre schème d'annotation. Deux développements notables interviendront par la suite, l'un d'eux pour permettre de combiner différentes segmentations du discours (analyses linguistiques en complément à nos analyses lexicales), l'autre consistant à prendre en compte différents algorithmes de structuration du discours afin de comparer plusieurs modes de parcours de textes.

Sur un plan plus fondamental, notre approche du discours suivant différents angles d'observation thématique révèle une faiblesse théorique de l'analyse thématique du discours. C'est dans cet axe de recherche que nous dirigerons nos prochains travaux.

Remerciement

Nous remercions Haifa Zargayouna, Guillaume Pitel, Olivier Ferret, Jean-François Condotta, et Vincent Barbier, pour leurs discussions et conseils ainsi que nos trois relecteurs anonymes pour la richesse de leurs commentaires.

Annexe : le vin jaune

En 1991, à la Station INRA de Dijon, Patrick Étiévant et Bruno Martin commençaient l'analyse du vin jaune, produit seulement dans le Jura. Le goût spécifique de ces vins résulte de leur technique d'élevage : on laisse le vin vieillir en tonneau pendant plusieurs années, sous un voile épais de levures Saccharomyces cerevisiae. Ce type de vin est également fabriqué en Alsace, en Bourgogne et à Gaillac sous le nom de vin de fleur ou vin de voile ; il n'a d'équivalent à l'étranger que dans le xérès, les sherrys ou le tokay de Hongrie. Quelles molécules sont responsables de son goût caractéristique ?

Les vins contiennent des centaines de composés volatils, dont un dixième sont aromatiques, de sorte que la détection des molécules responsables d'un arôme particulier est notoirement difficile : chercher le coupable, parmi 300 suspects... Au début des années 1970, certains avaient cru que la solérone (le 4 acétyl gamma butyrolactone) était l'arôme principal du vin jaune, mais, en 1982, Pierre Dubois, à Dijon, retrouva la solérone dans des vins rouges : la molécule avait un alibi.
On soupçonna alors le 4,5 diméthyl 3 hydroxy 2(5H) furanone, ou sotolon, molécule construite autour d'un cycle de quatre atomes de carbone et d'un atome d'oxygène. Comme le sotolon et la solérone sont en concentrations minimes dans les vins de voile et, de surcroît, chimiquement instables, les chimistes dijonnais ont cherché à optimiser leur extraction afin de déterminer la molécule responsable du goût de jaune.

L'analyse la plus directe d'extraits de vins est la chromatographie : on injecte un échantillon dans un solvant que l'on vaporise et on fait traverser au mélange une colonne revêtue intérieurement d'un polymère, qui retient les divers composés du mélange à des degrés divers ; en bas de la colonne, on détecte la sortie des composés séparés. Le premier travail des chimistes fut la mise au point d'une variante de cette technique pour identifier les composés présents en quantités minimes dans des mélanges complexes.
Les chromatogrammes d'échantillons de vin furent alors comparés à ceux de solutions pures de sotolon et de solérone de synthèse : le sotolon est ainsi présent entre 40 et 150 parties par milliard dans les sherrys ; la solérone semble moins spécifique, et ses concentrations sont supérieures dans les sherrys, ce qui explique pourquoi on l'a d'abord trouvée dans ces vins.

Enfin les dosages, complétés de tests sensoriels des fractions séparées, montrèrent que la solérone, aux concentrations trouvées dans du savagnin (le cépage à partir du quel on fabrique le vin jaune), n'était perçue par les consommateurs ni dans les vins, ni dans des solutions modèles : la solérone n'était pas la molécule caractéristique ; le jugement était sans appel.

En 1992, les chimistes se consacrèrent alors complètement au sotolon, qui avait été observé dans des molasses de canne à sucre, dans des graines de fenugrec, dans de la sauce de soja, dans du saké... Il était également présent dans certains vins botrytisés, c'est-à-dire faits à partir de raisins surmaturés et atteints par la pourriture noble : ce champignon, Botrytis cinerea, fait, par exemple, les sauternes ou les vins dits de vendanges tardives. Le sotolon n'a pas été trouvé dans les vins rouges ni dans les vins oxydés et, surtout, il fut déterminé que son seuil de perception était de 15 parties par milliard seulement.

Mieux encore, des tests de consommation montrèrent que les vins de voile étaient jugés typiques, avec une note de noix, quand la concentration en sotolon était forte dans ces vins. A plus forte concentration, les jurys de dégustation décrivirent une note de curry.

La piste du sotolon est aujourd'hui suivie par Élisabeth Guichard, qui a mis au point une méthode rapide de dosage : la concentration en sotolon dans le vin de paille (un vin préparé à partir de baies séchées sur des claies), qui n'avait pas été observée, est comprise entre 6 et 15 parties par milliard; le sotolon du vin jaune est synthétisé à la fin de la phase de croissance exponentielle des levures. Dans des vins vieillis respectivement un an, deux ans, trois ans, quatre ans, cinq ans et six ans, la quantité de sotolon est faible dans les débuts de la maturation et augmente notablement après quatre ans d'élevage, surtout dans les caves pas trop fraîches.

Des prélèvements à différentes profondeurs, sous le voile, dans les tonneaux, ont révélé que le sotolon est deux fois plus concentré au milieu et au fond des tonneaux que juste sous le voile. On suppose que le sotolon est indirectement produit par les levures du voile, quand le degré alcoolique est élevé : celles-ci transformeraient un acide aminé du vin en un cétoacide, qui serait libéré à la mort des levures, tombant au fond du tonneau ; puis une réaction chimique transformerait le cétoacide en sotolon, enrichissant d'abord le fond, puis le milieu, puis les couches supérieures du vin.

Puisque le sotolon est bien la molécule du goût de jaune, on cherche aujourd'hui des souches de levures qui ont la capacité d'en produire beaucoup ; on cherche aussi les conditions qui favorisent la production de ce goût.

Bibliographie

(Asher and Lascarides, 1994) N. Asher and A. Lascarides, “Intentions and information in discourse”, *Proceedings of the 32nd ACL*, pp. 34-41, 1994.

(Ben Hazez and al., 2001) S. Ben Hazez, J.-P. Desclés et J.-L. Minel, “ Modèle d’exploration contextuelle pour l’analyse sémantique des textes”, TALN, pp. 73-82, Tours, 2001.

(Boguraev and al., 1999) B. Boguraev, R. K. E. Bellamy, C. Kennedy, “Dynamic Presentation of Phrasally-based Document Abstractions”, *HICSS*, 1999.

(Boguraev and al., 1997) B. Boguraev and C. Kennedy, “Salience-based content characterisation of text documents”, *In Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, pp. 2-9, Madrid, Spain, 1997.

(Bird and al., 1999) S. Bird and M. Liberman, “A Formal framework for linguistic annotation”, *Technical Report* MS-CIS-99-01, Deparment of Computer and Information Science, University of Pennsylvania, 1999.

(Cristea and al., 1998) D. Cristea, N. Ide and L. Romary, “Marking-up views of a Text: Discourse and Reference”, *Proceedings of LREC*, 1998.

(De Rose and al., 2001) S. De Rose, E. Maler, D. Orchard, “XML Linking Language (Xlink) V1.0.”, W3C Recommandation, 27 June 2001.

(Dybkjær and al., 2000) L. Dybkjær and N. O. Bernsen, “The MATE Markup Framework”. In Dybkjær, L., Hasida, K. and Traum, D. (Eds.): *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue*, Hong Kong, October 2000.

(Ferret and al., 1998) O. Ferret, B. Grau et N. Masson, “Thematic segmentation of texts: two methods for two kinds of texts”, *Actes ACL-COLING'98*, Montréal, Canada, volume 1, pp. 392-396, 1998.

(Ferret and Grau, 1998) O. Ferret et B. Grau, “A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts”, *ECAI*, pp. 155-159, 1998.

(Grishman, 1998) R. Grishman, “TIPSTER Text Architecture Design”, Version 3.1, New York University, 7 October 1998.

(Grosz and Sidner, 1986) B. Grosz and C. Sidner, “Attention, Intentions and the structure of discourse”, *Computational Linguistics*, 12(3), pp. 175-204, 1986.

(Hearst, 1997) M. A. Hearst, “TextTiling: Segmenting Text into Multi-paragraphe Subtopic Passages”, *Computational Linguistics, 23*, 1, pp. 33-64, 1997.

(Hobbs and al., 1993) J.R. Hobbs, M. Stickel, D. Appelt and P. Martin, “Interpretation as abduction”, *Artificial Intelligence*, vol. 63, pp. 69-142, 1993.

(Hovy, 2001) E. Hovy, “The Future of Summarization”, Workshop on Text Summarization. New Orleans, Louisiana USA, September 2001.

(Ide and al., 2000) N. Ide, P. Bonhomme, L. Romary, “XCES: An XML-based Standard for Linguistic Corpora”, *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 825-30, 2000.

(Kozima, 1993) K. Hideki, “Text Segmentation Based on Similarity between Words”, *In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics* (Student Session), Colombus, Ohio, USA, 1993.

(Lin and Hovy, 1997) C-Y Lin and E. Hovy, “Identify Topics by Position”, *Proceedings of the 5th Conference on Applied Natural Language Processing*, March 1997.

(Lopez and Romary, 2000) P. Lopez and L. Romary, “A Framework for Multilevel linguistic Annotations*”, LREC Workshop on Meta-Descriptions, and Annotation Schemes for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora*, Athens, 2000.

(Maler and De Rose, 1998) E. Maler and S. De Rose, “XML Pointer Language (XPointer)”, *W3C Working Draft*, 3 March 1998 (see also http://www.w3.org/TR/1998/WD-xptr).

(Mani and Maybury., 1999) I. Mani and M. T. Maybury, “Advances in automatic text summarization”, *MIT Press*, Cambridge, MA, 1999.

(Mann and Thompson, 1988) W. C. Mann and S. A. Thompson, “Rhetorical Structure theory: toward a functional theory of text organization”, Text, 8(3), pp. 243-281, 1988.

(Marcu, 2000) D. Marcu, “Rhetorical Parsing of Unrestricted Texts”, *Computational Linguistics*, vol. 26, n°3, pp. 395-448, 2000.

(Masson, 1995) N. Masson, “An automatic method for document structuring”, *Actes 18th ACM-SIGIR* , Seattle, USA, pp. 372-373, 1995.

(Masson, 1995) N. Masson, “ Méthodes pour une génération variable de résumé automatique : vers un système de réduction de texte”, Thèse de doctorat, Université Paris XI, 1998.

(Mather and Note, 2000) L. A. Mather and J. Note, “Discovering Encyclopedic Structure and Topics in Text”. *Sixth ACM SIGKDD*. Boston, MA, USA, August 2000,.

(Porhiel, 2001) P. Sylvie, “Linguistic expressions as a tool to extract thematic information”, *Actes Corpus Linguistic*, Lancaster University, 2001.

(Salton and al., 1996) G. Salton, A. Singhal, C. Buckley and M. Mitra, “Automatic Text Decomposition Using Text Segments and Text Themes”, *Actes Hypertext'96, Seventh ACM Conference on Hypertext*, Washington, D.C., pp. 53-65, 1996.

(Sperberg-McQueen and Burnard, 1994) C. M. Sperberg-McQueen and L. Burnard, “Guidelines for Electronic Text Encoding and Interchange”, *ACH-ACL-ALLC Text Encoding Initiative*, Chicago and Oxford, 1994.

(Yaari, 1997) Y. Yaari, “Segmentation of Expository Texts by Hierarchical Agglomerative Clustering”, *Proceedings of RANLP* ,1997.