UniTHEM, un exemple de traitement linguistique à couverture multilingue

Nadine Lucas et Emmanuel Giguet

GREYC, CNRS UMR 6072, Université de Caen, 14032 Caen Cedex

Nadine.Lucas@info.unicaen.fr
Emmanuel.Giguet@info.unicaen.fr

Résumé :

Un logiciel d'analyse thématique à couverture multilingue est présenté. Le programme prend en entrée un texte HTML et renvoie en sortie le texte colorié en fonction des thèmes traités, en proposant une vue de la hiérarchie des sous-thèmes. Ce logiciel appelé UniTHEM accepte des langues à écriture alphabétique (langues latines, anglais, … russe) mais aussi les écritures à graphie liée (chinois, japonais). Les limites actuelles de couverture tiennent à des particularités de format d'une part, à la longueur du texte d'autre part. En effet, les textes structurés par des intertitres ne sont pas analysés comme tels. Ces limites montrent que la démarche n'est pas statistique ni basée sur des mots-clés. Elle s'appuie sur un modèle théorique de l'exposition, mis en relation avec des traits stylistiques, ce qui permet l'exploitation de la mise en forme matérielle du document, qui est relativement invariante. Les indices exploités sont communs à des familles d'écriture. Les ressources sont limitées aux séparateurs graphiques. Ces données permettent de constituer une hiérarchie des unités thématiques traitées par recoupements successifs des contextes. La qualité des analyses obtenues est satisfaisante. Les problèmes relatifs à l'évaluation de tels outils sont évoqués.

MOTS-CLÉS : recherche d'information, documents multilingues, analyse de texte, mise en forme matérielle, TAL robuste, thématique, Unicode.

Abstract :

This paper introduces a language-free topic parser. The task is to highlight the theme-topic structure and the hierarchy of subtopics in a text. It is performed on newspapers and magazines in French, English and various European languages, then extended to

different writing systems, such as Cyrillic and Chinese. Resources are common separators, punctuation and repeated segments. The algorithm relies on a linguistic model that allows to link stylistic features to the topic structure. The layout of text provides information on the stylistic features.

KEY-WORDS : robust text parsing, cross language information retrieval, layout retrieval, topic subtopic hierarchy, Unicode.

1. Introduction

La recherche des "thèmes" traités dans un document est une tâche ordinairement basée sur un traitement statistique des mots-clé censés représenter un concept (ou de termes représentant un concept) [Segond 2002 ; Toussaint 2004]. Ceci présuppose que le mot est une unité universelle [Salton 1989 ; Grefenstette 1998]. Cette vision occidentale domine largement le domaine de la recherche d'information. Cependant, nombre de "langues" ou d'écritures n'utilisent pas le mot graphique, entre autres, le chinois et la majorité des écritures d'Asie, ainsi que l'arabe. On peut s'étonner que la vision dominante ne soit pas davantage contestée par des auteurs issus de traditions scripturales différentes, mais elle l'est parfois (poliment) [Ogawa 1995; Chen 1997 ; He 2002].

L'approche ici présentée s'écarte de la vision dominante en informatique en ce que le mot n'est pas posé comme pivot ou atome de sens. L'unité thématique dans un texte n'est pas caractérisée lexicalement, mais plutôt spatialement comme un passage de texte traitant (en principe) d'une même idée. Cette conception est assez proche de la vision de Hearst connue sous le nom de *text tiling* [Hearst 1994, 1997]. Mais, contrairement à celle de Hearst, notre approche est fondée sur un modèle de linguiste. Quoique le résultat soit visuellement similaire, les présupposés sont différents. Le modèle de référence est celui de Yamada (1873-1958), un linguiste japonais qui définit des opérations de mise en discours sans s'appuyer sur le concept de mot (lequel n'est pas matérialisé en japonais). Il présente des opérations qui peuvent être exprimées comme une série de contraintes qui limitent *de facto* un exposé qui serait sinon un développement infini. On va donc délimiter des passages de texte, comme ayant une cohésion interne, mais il faudra aussi hiérarchiser ces passages, puisqu'il est possible qu'un exposé contienne des développements subordonnés ou incis, qui interrompent momentanément la progression du discours. L'intérêt principal d'une telle approche basée sur des considérations stylistiques est qu'elle fait appel à la perception (de la mise en forme du texte). Elle a une portée générale et s'applique sans modification profonde à des textes en chinois ou dans une autre écriture qui ne connaît pas le mot graphique, mais aussi aux langues occidentales.

Du point de vue informatique, notre travail se situe dans le paradigme des analyseurs robustes, à ressources limitées et à couverture multilingue. Considérant le défi du multilinguisme, nous avons recensé les techniques pouvant s'appliquer à des textes non caractérisés en langue. La présente étude s'appuie sur les traitements sans dictionnaire [Vergne 2001], certains concepts comme l'apprentissage robuste générant de nouvelles connaissances (techniques dites d'induction) [Kushmerick 1999] ou la déduction contextuelle [Muslea *et al.* 2002a, b], techniques

ordinairement appliquées à l'exploration de la toile, que nous exploitons aussi dans l'analyse des textes. Quoique l'objectif soit celui de la recherche d'information multilingue (*cross language information retrieval*) [Oard 1997 ; Grefenstette 1998; Collins & Singer 1999], nous ne nous focalisons pas sur la recherche de mots ni d'entités nommées. Nous extrayons plutôt les connecteurs [Déjean 2000].

Partant d'un logiciel d'étude destiné à détecter et à colorier les « unités thématiques » dans des articles en français, nous nous sommes attachés ici à tester les possibilités d'adaptation de ce logiciel à un corpus multilingue et même « multi-script », c'est-à-dire traitant des documents non alphabétiques et sans mot graphique. Nous présentons dans la section 2 le logiciel THEMA dans un contexte français, puis ses avatars dans la section 3. Le premier, EuroTHEM, est destiné à des textes européens (Iso-latin), ce qui ne nécessite pas de modifications profondes du logiciel. Le second avatar, UniTHEM, est adapté à des textes que l'on peut traiter sous Unicode, incluant donc les textes en graphie liée, informatiquement gérables sous utf-8. Nous discutons dans la dernière section des caractéristiques permettant la re-définition du multilinguisme dans la perspective du TAL appliqué aux documents.

2. Le fil du discours selon THEMA

2.1. Principes généraux

Le logiciel THEMA est à l'origine un logiciel d'étude, développé pour détecter des unités thématiques. L'objectif de ce travail était de mettre en œuvre une stratégie d'analyse textuelle à l'échelle du document dans des articles courts en français [Pinatel, 2003]. Cette analyse est basée sur des critères rhétoriques et stylistiques, en cela elle ressemble à celles de Kando et Karlgren [Kando 1997 ; Karlgren 2000]. Mais contrairement à ces travaux, nous avons cherché à limiter la dépendance au lexique, qui est un obstacle à la généricité. Le modèle de référence que nous avons choisi est celui de Yamada [Yamada 1936], linguiste et sémioticien souvent évoqué par des chercheurs japonais en RI [Hirakawa 1989 ; Sakamoto & *al*. 2002]. Ce modèle met l'accent sur les procédés d'exposition dans le discours. Il met en valeur le fil du discours, et pour cela traite de divers échelons de la hiérarchie des thèmes (l'enchâssement de thèmes subordonnés) et des unités repérées par rapport au titre, les thèmes dérivés notamment. Pour la première étape, le logiciel devait traiter d'un corpus de textes de vulgarisation en français, donc de textes d'abord facile.

Ce logiciel présente un certain nombre de défauts et de limitations. Cependant, il est suffisamment robuste pour produire des analyses correctes dans la majorité des cas et supporter la comparaison avec des logiciels plus classiques de détection des thèmes [Hernandez & Grau 2002]. Les ressources linguistiques utilisées sont légères. Pour le cas où une marque attendue n'est pas détectée dans le texte entrant, des procédures d'induction, exploitant les ressources endogènes (présentes dans le texte à analyser) sont mises en œuvre, suivant la méthode distributionnelle [Déjean 2002]. Ces propriétés permettent d'envisager une couverture multilingue. Nous avons cherché à tirer parti de la norme ISO-CEI 10646, couramment appelée Unicode, une

révolution technologique qui permet d'aborder le traitement informatique de l'écrit avec un degré d'abstraction suffisant [Andries 2002].

2.2. Description du logiciel de base

Les documents fournis au format HTML sont segmentés. La mise en forme matérielle ou MFM [Virbel & Pascual 1996] a une grande importance, puisque le titre et le corps de texte doivent être correctement délimités. Il est nécessaire d'initialiser le traitement en extrayant le titre. Nous retenons comme informations pertinentes la position et la mise en forme différentielle d'une sous-chaîne de caractères, qui caractérisent le titre de l'article. Par différentielle, on entend tout simplement une MFM différente du reste de l'article. De même, la position est remarquable, le titre étant soit le premier élément de l'article, soit un élément isolé placé dans le premier tableau. Ce segment (la chaîne de caractères qui est censée contenir le titre) est posé comme thème de niveau 1 ou G pour global.

Notons toutefois que la segmentation typographique n'est qu'une étape. Le logiciel peut renvoyer comme résultat que le titre et le chapeau d'un article forment le segment thématique au niveau global. Dans la version ici présentée de THEMA, la segmentation interne au corps de texte est restée primitive, autrement dit les textes sont segmentés en unités typographiques fixes : paragraphes, phrases et virgulots (unités délimitées par une virgule). Une version ultérieure du logiciel traite des unités typographiques telles que les sections et chapitres, mais pour l'expérience présente nous nous centrons sur l'aspect multilingue et traitons de textes journalistiques courts.

La segmentation du texte est faite par le module *TextTokenizer*, en unités fixes, paragraphe, phrase et virgulot. Ce module permet également de reconnaître et de garder en mémoire la mise en forme matérielle ainsi que la hiérarchie des composants. Leurs attributs éventuels de MFM (en-ligne) tels que graisse, italique, couleur, sont également mémorisés ainsi que les caractéristiques d'alignement.

La segmentation en paragraphe est une segmentation peu dépendante des familles d'écriture, c'est une délimitation du document HTML. Les codes HTML mal formés forment un obstacle à l'analyse, mais le programme Tidy[1] de Dave Raggett permet d'en corriger un grand nombre. La segmentation en paragraphe peut être modifiée dans une fenêtre interactive, les paragraphes que l'on ne considère pas comme faisant partie du corps de texte (mentions éditoriales en particulier) peuvent ainsi être décochés et exclus du traitement ultérieur. Cette fonctionnalité n'est pas utilisée dans les exemples proposés, sauf mention contraire. On verra donc le résultat automatique.

Les opérations sous-jacentes au modèle (mise en facteur d'un thème dans une unité thématique) sont implémentées dans un algorithme de mise en relation correspondant au modèle de Yamada, baptisé *Thematisation*. Nous ne nous attarderons pas ici sur les principes de base de l'algorithme. La structure des exposés est représentée par une série de contraintes sur un développement infini.

[1] http ://www.w3.org/People/Raggett/tidy.

Les ressources en mémoire sont les ponctuations et une cinquantaine d'items qui forment une base d'indices dits morphologiques stockés dans une base de données MySQL. Nous avons en effet établi pour le français les marques spécialisées pour la détection d'exposés (marques d'ouverture et clôture de thème, marques de subordination et disjonction). Ces marques sont hiérarchisées. En effet, si l'on trouve deux phrases coordonnées par *De plus*, cela n'a pas la même valeur que deux phrases coordonnées par *Et*. Deux paragraphes coordonnés par *De plus* n'ont pas le même statut que deux phrases coordonnées par ce connecteur, même si en français, la marque de coordination est la même.

Le logiciel ne comporte pas de diagnostic de langue puisqu'il est censé travailler sur le français. Il est important de noter que les ressources en mémoire sont exploitées lorsqu'elles sont présentes dans le texte entrant, mais qu'elles ne sont pas la seule source de connaissances. La déduction contextuelle informée par le modèle, ou si l'on veut, l'induction des marques de bornage (*wrapper induction*) est implémentée.

2.3. La détection des thèmes

Les unités thématiques sont calculées et présentées sur trois niveaux d'inclusion, autrement dit, pour un article normal le premier niveau G subdivise l'article en titre et corps de texte. On suppose ici que le titre est en fonction thématique détachée, et que le corps de texte est le développement (ou propos) qui apporte de l'information (au niveau global). Le corps de texte est subdivisé ensuite en sous-thèmes au niveau G1 et si besoin est, également à un troisième niveau G11 (si le texte est long ou dense). A ces trois niveaux successifs, les informations pertinentes sont analysées automatiquement, à partir d'indices graphiques (MFM), morphologiques (les marques), d'indices positionnels (début et fin) et d'indices de niveau (trois échelons correspondant aux trois grains ou mesures de texte).

Thème G	**Coraux** en mal de **squelette**	
	Thème G1	"Nous avons montré, pour la première fois que, outre son **rôle** dans le réchauffement global, **l'augmentation** de la **concentration** de gaz carbonique (CO2) dans l'atmosphère a un effet négatif sur un écosystème **marin** : les **récifs** coralliens" explique **Jean-Pierre Gattuso**, chargé de recherche au laboratoire Océanographie biologique et écologie du plancton marin de Villefranche-sur-Mer.
	Thème G11	Comment? En réduisant, tout simplement, la possibilité pour les coraux de se constituer un squelette calcaire **à** partir de calcium et de carbonate (CO3). En effet, l'augmentation de CO2 provoque une diminution de la concentration en CO3 néfaste au **processus** de **calcification** : "Nous estimons que celle-ci a diminué de 8% depuis 1880 et qu'elle pourrait encore baisser de 20% d'ici 2065" poursuit le chercheur.

Figure 1. Le fil du discours déroulé sur trois niveaux d'inclusion

La répétition de certaines sous-chaînes (en gras dans l'exemple ci-dessus) constitue un critère, mais il est important de pouvoir situer des répétitions par rapport à une mesure de texte. En ce sens, les répétitions ne sont pas à entendre comme dans les techniques tf-idf, mais dans une acception stylistique (anaphore et

épiphore). La démarche consistant à exploiter des indices morphologiques et des positions est commune aux systèmes de type syntaxique de complexité réduite [Vergne 2001]. Lorsque les indices morphologiques sont absents, l'algorithme emprunte aux systèmes d'apprentissage de règles grammaticales [Déjean 2002]. Cet apprentissage n'est pas itératif sur un corpus, il ressemble donc davantage aux principes dits d'induction, utilisé pour le dépouillement de sites.

Les résultats sont proposés sous plusieurs formes. La « structure compacte » permet de donner une vue globale du texte avec les débuts des unités thématiques, dont le développement est évidé si le texte est long. Les unités thématiques de rang 1, 2 et 3 sont ensuite présentées séparément, pour permettre la vérification de pertinence des relations hiérarchiques. Les unités de rang 1 sont plus longues et moins nombreuses, et on procède ainsi par raffinement successif jusqu'aux plus petites structures détectées qui sont incluses dans les développements. La structure développée représente les unités thématiques et leurs inclusions pour l'ensemble du texte. Nous présentons ici les structures abrégées dites structures compactes.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1

THEME	Coraux en mal de squelette
RHEME	*Unité Thématique G1 , niveau 2*
	THEME "Nous avons montré, [...] chargé de recherche au laboratoire Océanographie biologique et écologie du plancton marin de Villefranche-sur-Mer.
	RHEME *Unité Thématique G11 , niveau 3*
	THEME Comment? [...] "Nous estimons que celle-ci a diminué de 8% depuis 1880 et qu'elle pourrait encore baisser de 20% d'ici 2065" poursuit le chercheur.
	RHEME Or, [...] PEDRO LIMA

Figure 2. Résultat de la structuration thématique en français

On y voit que le thème le plus général G (le titre, en bleu clair) est informé par le corps de texte (en vert clair), qui comprend un seul thème de niveau 2 correspondant à une reformulation plus détaillée (bleu plus soutenu), le rhème étant constitué également d'une unité thématique. On voit ici que le dégradé des thèmes n'est pas accompagné par un dégradé des rhèmes, dans un texte monothématique. Autrement dit, les segments rhématiques et les segments de clôture restent tous au niveau 3 (vert le plus foncé). Le nom d'auteur n'est pas ramené au niveau du titre. Il s'agit d'une maladresse dans l'algorithme et dans la présentation des résultats. Mais pas seulement.

En effet, le modèle de Yamada n'est pas un emboîtement en poupées russes. Il permet en revanche une distinction entre un texte monothématique (exemple 2) et un texte plurithématique comme dans l'exemple 3 où le décrochement G12 signale l'articulation entre le thème 1 et le thème 2. Notre parti-pris dans THEMA a été de mettre en valeur le fil du discours. Une comparaison est proposée avec un autre modèle d'analyse dans la section 4.

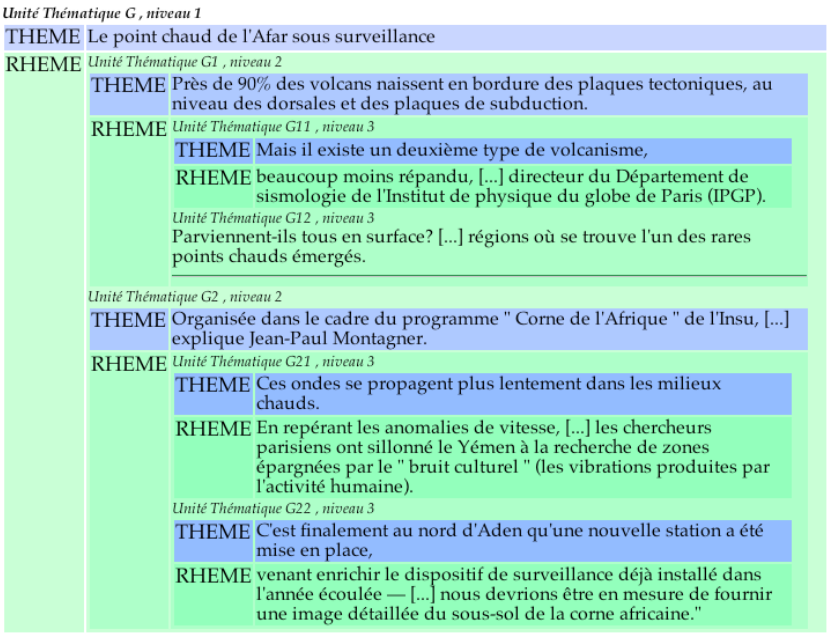

Figure 3. Résultat de la structuration thématique en français

3. Version « multilingue »

3.1. Langues latines et EuroTHEM

La première extension de THEMA a été la couverture aux langues latines (espagnol, portugais, italien, roumain), ce qui ne nécessitait pas de modification du point de vue informatique. En effet, les tests de détection de connecteurs (établis pour le français) ne sont pas conçus comme décisifs. La variation stylistique est assez grande dans le corpus traité (en français) pour que les formes attendues ne soient pas « obligatoires ». Comme on l'a vu, les traitements par défaut exploitent la répétition d'une forme *x* inconnue à une position *y* donnée, c'est donc le procédé stylistique qui est capté et exploité. Cette recherche est contrainte par l'examen des positions remarquables, en position de préfixe (au début d'une unité typographique) ou alors en position de suffixe (en fin d'une unité typographique). On spécifie alors un petit espace de recherche et on recense les chaînes répétées, par exemple dans la dépêche en italien (fig.4), *partiti*.

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1

THEME	Blitz dei partiti: pioggia di soldi in arrivo	
RHEME	*Unité Thématique G1 , niveau 2*	
	THEME	ROMA - Pioggia di miliardi pubblici sui partiti.
	RHEME	*Unité Thématique G11 , niveau 3*
		THEME Ieri infatti,
		RHEME con un voto perfettamente bipartisan, [...] compiuto ieri in commissione affari costituzionali.
		Unité Thématique G12 , niveau 3
		THEME "E' stato votato in trenta secondi - racconta il parlamentare - all'unanimità.
		RHEME Mi sono astenuto solo io". [...] il provvedimento diventa immediatamente operativo senza dover passare per l'aula parlamentare.
	Unité Thématique G2 , niveau 2	
	THEME	Secondo il racconto di Boato, [...] mentre si discutevano questioni di importanza secondaria.
	RHEME	*Unité Thématique G21 , niveau 3*
		THEME In poco tempo,
		RHEME il provvedimento è stato presentato e votato, [...] e che il rimborso per ogni voto passa da due a cinque euro.
		Unité Thématique G22 , niveau 3
		THEME Aumenta di cinque volta anche il rimborso per le elezioni regionali,
		RHEME che sale da cinque centesimi a dieci.

Figure 4. Résultat de la structuration thématique en italien

3.2. Ecritures alphabétiques

Pour étendre la couverture de THEMA aux langues à écriture latine, nous avons dû séparer la détection des paragraphes et celle des unités ponctuées (phrase, virgulot). En effet, la structuration en paragraphe est invariante, tandis que la segmentation en phrases est dépendante d'une graphie liée à une famille de langues. En contexte multilingue, les difficultés de segmentation en phrases tiennent essentiellement aux variations dans les abréviations et dans l'utilisation de la capitalisation [Mikheev 2002; Kiss et Strunk 2002, 2004]. L'algorithme proposé par ces auteurs n'est pas implémenté dans notre logiciel, mais cette amélioration est envisagée. La segmentation effectuée par le module "texttokenizer" permet cependant de traiter des textes en anglais, allemand, néerlandais, langues slaves, ainsi que dans les langues latines.

Dans l'étape d'analyse, le logiciel exploite comme précédemment les répétitions pour affecter une valeur de préfixe ou suffixe à des chaînes de caractères. Cette procédure à fondement stylistique permet de délimiter des unités thématiques à partir de ressources strictement endogènes (contenues dans le texte).

STRUCTURE THEMATIQUE COMPACTE

Unité Thématique G , niveau 1

THEME	Robots to Gain Eyes in the Back of Their Heads
RHEME	*Unité Thématique G1 , niveau 2*
	THEME: LONDON (Reuters) - Researchers in the United States are developing robots with "eyes in the backs of their heads" in the form of nine digital cameras attached to a frame the size of a beach ball. [...] A report on their work is in the latest edition of the New Scientist magazine.
	RHEME: *Unité Thématique G11 , niveau 3*
	THEME: The new "eye, [...] many robots have to rely solely on their single eye.
	RHEME: But as computer scientists at the University of Maryland proved mathematically in 1998, [...] so they were less likely to fail or disappoint.

Figure 5. Résultat de la structuration thématique en anglais

Le russe, plus exactement l'écriture cyrillique, qui détache les mots, ne pose pas de problèmes particuliers du point de vue informatique. Les textes sont segmentés en paragraphes, puis en phrases et en virgulots. Les codes des glyphes correspondant au point de fin de phrase et à la virgule sont les mêmes qu'en Iso-latin.

Unité Thématique G , niveau 1

THEME	Колокольный звон над колонией
RHEME	*Unité Thématique G1 , niveau 2*
	THEME: раздастся скоро - здесь будет построена церковь [...] не опускают руки поборники доброты и нравственности.
	RHEME: *Unité Thématique G11 , niveau 3*
	THEME: Беспрецедентное событие не только для Зеленограда,
	RHEME: Москвы, [...] которая станет Подворьем Данилова мужского монастыря Москвы.
	Unité Thématique G12 , niveau 3
	THEME: Церемонию закладки в фундамент церкви капсулы с грамотой на освящение проводил архимандрит Алексий, [...] взявшие на себя обязательства по финансированию строительства храма.
	RHEME: Перед Богом все равны, [...] истинное.
	Unité Thématique G2 , niveau 2
	THEME: - Прихожане здесь отличаются от обычных искренностью, [...] стал священнослужителем и ведет приход в Подмосковье.
	RHEME: *Unité Thématique G21 , niveau 3*
	THEME: ...Когда наступил момент торжественной службы, [...] что права гражданина соблюдаются.
	RHEME: Церкви есть практически в каждой колонии, [...] Л.РОМАНОВА

Figure 6. Résultat de la structuration thématique en russe

3.3. Version multi-script UniTHEM

Le traitement des caractères en utf-8 est à ce stade indispensable. La détection automatique du jeu de caractères entrant est suivie d'une conversion automatique vers utf-8. Les entités HTML sont également converties vers utf-8. L'ajout de cette fonction dans UniTHEM a pour conséquence un ralentissement de la procédure de segmentation, par rapport à la version en Iso latin, les fichiers sont aussi plus lourds. Le traitement des caractères en utf-8 et les classes d'équivalence de ponctuation nous permet d'aborder des écritures dites à graphie liée, pour des articles en chinois et japonais, qui utilisent des idéogrammes et ne séparent pas les mots. Le coréen, qui ne sépare pas les mots mais utilise un syllabaire n'est pas présenté ici faute de place. Les principes sont les mêmes, mais il est nécessaire de définir des classes de ponctuations [Giguet *& al*. 2000].

Les documents en japonais et chinois sont segmentés en paragraphes, puis en phrases et en virgulots, ce qui nécessite de prendre en compte les codes des glyphes correspondant au point de fin de phrase et à la virgule. On prend également en compte les ponctuations spécifiques de ces écritures (point de mot équivalent au trait d'union). On notera que la segmentation en mots est tout simplement omise, car elle n'est pas indispensable à l'analyse. Contrairement aux traitements lexicographiques, basés sur la consultation de dictionnaires, notre approche exploite des chaînes de caractères répétés à certaines positions et la disposition d'ensemble du texte.

De la même manière que précédemment, nous retenons comme informations pertinentes la position et la mise en forme différentielle de segments de textes ou sous-chaînes, qui délimitent le titre de l'article. Ce segment est posé comme thème de niveau 1. Comme indiqué ci-dessus, dans l'étape d'analyse, le logiciel exploite les répétitions d'une suite de caractères (entre le titre et le corps de texte, et à l'intérieur du corps de texte) pour affecter une valeur à des segments bornés par des préfixes ou suffixes. Les résultats sont tout à fait corrects, avec les mêmes imperfections ou biais que dans les langues latines, mais aussi les mêmes atouts.

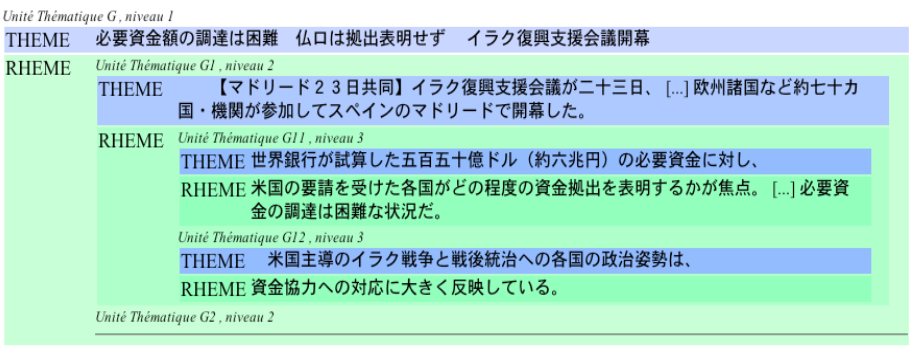

Figure 7. Résultat de la structuration thématique en japonais

La figure 7 montre que le thème le plus général G est informé par le corps de texte, lui-même subdivisé en un thème de niveau 2 (le 1er §) et un rhème explicatif

en deux parties (G11 objectifs et G12 commentaire). Les unités les plus petites au niveau 3 sont subdivisées en virgulots.

Dans la figure 8 on voit que le thème le plus général G (le titre, en bleu clair) est informé par le corps de texte (en vert clair), lui-même subdivisé en deux sous-thèmes (récit des faits et commentaire). Le dernier § est traité comme un ajout ou digression. La signature apparaît comme dernier rhème (les clôtures n'étant pas différenciées des rhèmes).

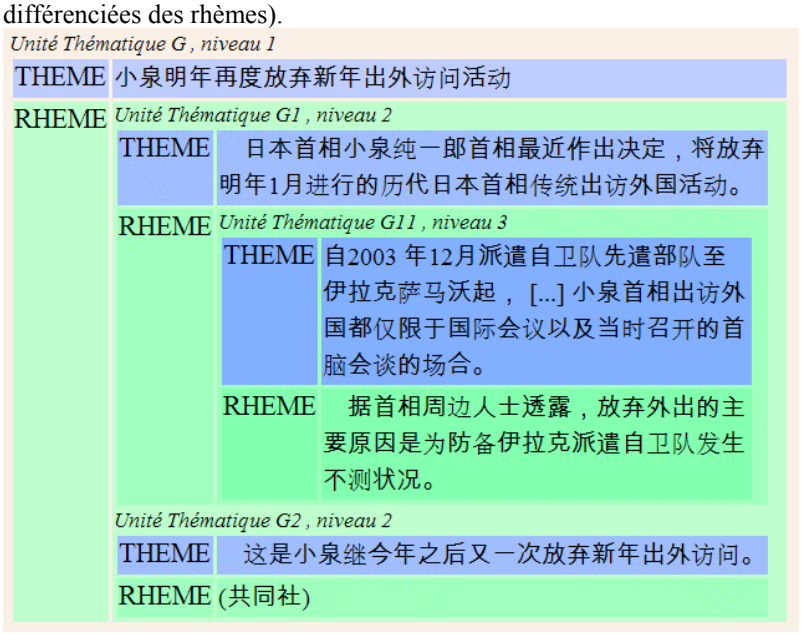

Figure 8. Résultat de la structuration thématique en chinois

Figure 9. Résultat de la structuration thématique en arabe

Les textes en arabe pris sur les sites de presse ne présentent pas une graphie liée. Ils ne nécessitent pas de traitement particulier, à part la gestion des indications de changement de sens d'écriture (*right to left mark*). La visualisation gère le calage à droite des paragraphes.

4. Résultats et évaluation

Le logiciel UniTHEM donne satisfaction pour le traitement des unités thématiques indépendamment des langues et des écritures. Les exemples cités ci-dessus proviennent d'internet et n'ont pas été retouchés manuellement par décochage des paragraphes. Les possibilités d'exploitation des documents sous Unicode nous semblent parfaitement mises en valeur. L'avantage d'UniTHEM est de présenter une vision synthétique du contenu des articles à travers la structure compacte. Cela permet de saisir l'information très rapidement, fournissant ainsi une aide appréciable à la lecture. Une nouvelle voie est ouverte pour aller au-delà des traitements purement statistiques et des traitements linguistiques lexicaux. Le logiciel a été intégré à la plate-forme « wims » [Giguet 2005] et il fonctionne en ligne.

La couverture des formats de documents est imparfaite. Nous sommes limités par le format d'entrée HTML et nous avons encore des difficultés avec certains documents que l'utilitaire Tidy ne permet pas de corriger.

L'évaluation qualitative de tels outils d'analyse thématique pose des problèmes, car le consensus entre différents courants ou experts n'est pas obtenu aussi facilement que sur l'analyse grammaticale par exemple. Notre algorithme étant déterministe, il y a toujours une solution. Cette solution est discutable. Le choix du modèle de référence peut être critiqué, d'autant qu'il n'est pas beaucoup vulgarisé.

Nous avons soumis des résultats d'analyse à des lecteurs externes (langue maternelle à tester) pour juger de leurs réactions sur des ensembles de 10 dépêches ou articles tirés aléatoirement (structure développée et structure compacte). La question posée était « l'analyse est-elle correcte, permet-elle de souligner le fil du discours ? ».

Langues /jugement	Nbre de textes	dont courts	non	oui	ne sait pas	nb oui/10
Chinois	10	5	3	5	2	5/10
Arabe	10	2	8	1	1	1/10
Allemand	10	7	2	8		8/10
Français	20	12	2	17	1	9/10
total	50	26	15	31	3	6/10

Tableau 1. Jugement des lecteurs sur la pertinence de l'analyse

Le ratio de « oui » indique un bon taux de satisfaction pour les textes courts. Une limitation patente est que l'analyse est faite sur trois niveaux seulement, ce qui est pénalisant pour les textes longs (plus de 10 §). Cela est net pour l'arabe, où la moyenne des paragraphes par article de notre échantillon avoisinait 20. Les hésitations s'observent sur les textes particuliers (interviews, éditoriaux). Enfin,

concernant la visualisation, nous envisageons une représentation des segments de clôture différente de celle des rhèmes profonds, car ce défaut suscite aussi des critiques.

Les commentaires d'auteurs d'articles en français rejoignent ceux des lecteurs, les textes longs étant invariablement jugés moins bien analysés que les textes courts. En complément, nous avons soumis les résultats d'analyse à un analyste de presse (3 articles analysés finement pour le roumain). Les commentaires sont positifs, ils discutent surtout du modèle pour l'organisation des unités incluses.

A titre de comparaison pour le traitement informatique, nous avons analysé un texte en français « Le vin jaune » déjà présenté par [Hernandez & Grau 2002]. Par rapport au résultat proposé par ces auteurs (voir annexe, structure incluse en gris), on remarque que la relation disjointe entre le premier et les deux derniers paragraphes (commentés comme introduction et conclusion) n'est pas représentée dans notre analyse. Contrairement au modèle en poupées russes, notre analyse favorise le suivi du fil du discours en partant du titre. La relation « introduction-conclusion » est constitutive de l'unité thématique, aussi le segment conclusif apparaît dans le rhème. Par ailleurs, dans notre analyse, le thème de niveau 2 englobe deux paragraphes, qui introduisent une séquence explicative. Le modèle invoqué par Hernandez & Grau, au contraire, favorise la relation question-réponse, et traite la réponse comme unité distincte.

Unité Thématique G , niveau 1

THEME	le vin jaune
RHEME	*Unité Thématique G1 , niveau 2*
	THEME: En 1991, [...] la molécule avait un alibi.
	RHEME: *Unité Thématique G11 , niveau 3*
	THEME: On soupçonna alors le 4, [...] molécule construite autour d'un cycle de quatre atomes de carbone et d'un atome d'oxygène.
	RHEME: Comme le sotolon et la solérone sont en concentrations minimes dans les vins de voile et, [...] 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.
	Unité Thématique G12 , niveau 3
	THEME: Les chromatogrammes d'échantillons de vin furent alors comparés à ceux de solutions pures de sotolon et de solérone de synthèse : [...] ce qui explique pourquoi on l'a d'abord trouvée dans ces vins.
	RHEME: Enfin les dosages, [...] puis les couches supérieures du vin.
	Unité Thématique G13 , niveau 3
	THEME: Puisque le sotolon est bien la molécule du goût de jaune,
	RHEME: 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.

Figure 9. Résultat de la structuration thématique sur *le vin jaune*

Les textes que nous avons soumis à l'analyse sont des dépêches de presse ou des articles de journaux et de magazines, des textes fortement structurés du point de vue thématique. La MFM est riche et diversifiée. L'algorithme délimite des unités thématiques à partir de ressources endogènes uniquement. Il serait donc possible de désactiver la consultation des indices morphologiques pour le français. Cependant, disposer d'indices a pour avantage une plus grande rapidité du traitement, car le calcul des répétitions est plus coûteux que la détection d'une forme spécifiée. Il

serait préférable de rajouter un diagnostic de langue et des ressources morphologiques spécialisées pour traiter d'un corpus stable du point de vue des langues.

En revanche, pour la couverture d'un événement particulier ou d'un jour J, comme l'a réalisé manuellement une équipe d'analystes [van Dijk 1988], il est intéressant de proposer une analyse thématique informatisée de la presse avec une couverture Unicode, sans restriction due au lexique et à la langue. Le suivi de l'actualité sur la toile est une application naturelle pour le modèle de Yamada, bien adapté au genre journalistique. L'efficacité de cette approche dans d'autres genres est à l'étude.

Le revers de la médaille est que le pré-traitement qui consiste à analyser automatiquement la mise en forme matérielle nécessite un soin méticuleux. Les documents sont formatés de façon très variable, or l'analyse stylistique s'appuie sur la reconnaissance des unités typographiques et de leur rang dans la hiérarchie textuelle.

5. Discussion

La voie d'approche stylistique est explorée notamment par Karlgren dans l'analyse de la presse [Karlgren 2000, Karlgren & Järvinen 2002]. Les principes informatiques de base retenus dans notre approche sont hérités de la recherche d'information sur la toile, une situation où la langue des sites est inconnue, les règles de formatage également [Stienne & Lucas 2003]. On a donc recours à une alternance de déduction et d'induction à partir d'indices endogènes, dont on cherche à établir le statut informationnel [Mukerjea 1998 ; Muslea & *al*. 2002a, 2002b]. Il est tentant de relier le traitement des sites et le traitement du contenu en utilisant les mêmes principes discriminatoires à partir de ressources endogènes. La source exogène de connaissance permettant de faire des choix est alors le type de sortie requis par l'utilisateur (choix de la grille d'analyse, du niveau de détail requis, etc.).

L'analyse de documents et la fouille de textes présentent aussi des convergences. La caractérisation différentielle de sites (classés en deux catégories qualitatives d'après leur contenu) a été menée en France suivant des principes structuralistes exploitant des faisceaux de traits [Valette 2004]. La valeur accordée aux indices matériels (ponctuations, couleurs) est dépendante du contexte et des différentiels observables. Pour notre objectif, l'affectation de valeur est dirigée par le modèle psycho-linguistique de Yamada. On pourrait dire que la classification des segments de textes est faite par UniTHEM en autant de classes que le modèle en requiert (au moins 2), et se raffine en autant de classes que le modèle le permet (24 si on se limite à trois niveaux d'inclusion).

Pour traiter une autre problématique, par exemple la détection de l'argumentation ou celle des explications, il serait nécessaire d'exploiter un autre modèle de référence et donc d'écrire un algorithme différent. Nous jugeons préférable d'exploiter un seul modèle de référence à la fois.

L'expérience de traitement thématique à couverture multilingue forme un jalon dans une recherche émergente, permettant l'interprétation de données textuelles

indépendamment des langues et des écritures (ou presque). L'homogénéité de la norme ISO-CEI 10646, qui utilise un jeu unifié de caractères, sert de socle technologique. Unicode permet de mener des traitements robustes sur un corpus multilingue ou indifférencié en langues, en s'appuyant sur quelques données internes au code ; les tables d'équivalence de glyphes, notamment les ponctuations, deviennent la « pierre de Rosette » du traitement pour UniTHEM. La généricité de l'approche tient à l'exploitation des balises et des graphies qui ont un usage plus large que le lexique d'une langue particulière. Nous souhaitons suivre à l'avenir cette voie d'exploration prometteuse.

Références bibliographiques

[Andries 2002] "Introduction à Unicode et à l'ISO 10646" P. Andries, *Document numérique* Vol. 6 n° 3-4. (2002) pp. 51-88.

[Chen 1997] "Chinese Text Retrieval without using a Dictionary". A. Chen & J. He, *SIGIR,* 1997.

[Collins et Singer 1999] "Unsupervised models for named entity classification" M. Collins and Y. Singer. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[Déjean 2000] "ALLiS: a Machine Learning System for Natural Language Learning" H. Déjean *Conference on Natural Language Learning* Lisbon. 2000.

[Déjean 2002] "Learning Rules and Their Exceptions" H. Déjean *Journal of Machine Learning Research* 2: 669-693 (2002).

[Giguet 2005] "Modélisation de l'activité expérimentale du chercheur en traitement des langues sur corpus multilingues" E. Giguet *Journée de l'ATALA* Articuler les traitements sur corpus, 12 février 2005.

[Giguet & *al.* 2000] "Document structure identification illustrated on news dispatches" E. Giguet, N. Lucas & G. Cousin *CicLing-2000* A. Gelbukh (ed.), Mexico, Instituto politécnico nacional, 2000 pp. 415-428.

[Grefenstette 1998] Cross-Language Information Retrieval G. Grefenstette (ed.) Kluwer, 1998.

[He 2002] Finding the Better Indexing units for Chinese Information Retrieval., H. He, P. He, J. Gao, & C. Huang *First SigHAN Workshop on Chinese Language Processing*, 2002.

[Hearst 1994] "Multi-Paragraph Segmentation of Expository Text", M. Hearst *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces,NM, June 1994.

[Hearst 1997] "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages" M. Hearst *Computational Linguistics*, 23:1, (1997) pp. 33-64.

[Hernandez & Grau 2002] "Analyse thématique du discours : segmentation, structuration, description et représentation" N. Hernandez & B. Grau, In *Cide 5*, Tunis (2002).

[Kando 1999] "Text Structure Analysis as a Tool to Make Retrieved Documents Usable" N. Kando *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages*, Taipei, Taiwan, Nov. 11-12, 1999, 126-132.

[Karlgren 2000] Stylistic Experiments for Information Retrieval., J. Karlgren PhD thesis, Stockholm, Université de Stockholm, 2000.

[Karlgren & Järvinen 2002] "Foreground and background text in retrieval" J. Karlgren . & T. Järvinen, In Karlgren, J., Gambäck, B. Kanerva, P. (eds) *Acquiring (and Using] Linguistic*

(and World) Knowledge for Information Access, Menlo Park, Spring 2002, The American Association of Artificial Intelligence.

[Kiss & Strunk 2002] "Viewing sentence boundary detection as collocation identification" *Proceedings of KONVENS 2002,* Saarbrücken, pp. 75-82.

[Kiss & Strunk 2003] "Multilingual Least Effort Sentence Boundary Detection" Tibor Kiss, Jan Strunk Under review.

[Kushmerick 2000] "Wrapper induction: Efficiency and Expressiveness", N. Kushmerick *Artificial Intelligence* 118, 2000.

[Mikheev 2002] "Periods, Capitalized Words, etc." Andrei Mikheev *Computation Linguistics* 28:3 (2002) pp. 289-318.

[Mukherjea 2000] "WTMS: A System for Collecting and Analysing Topic-Specific Web Information", S. Mukherjea. *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, Netherlands, May 15-19, 2000. http://www9.org/w9cdrom/293/293.html

[Muslea *et al.* 2002 a] "Adaptive view validation: a case study on wrapper induction" I. Muslea, S. Minton, C. Knoblock. *Proceedings19th ICML*, 2002.

[Muslea *et al.* 2002 b] "Active+ Semi-Supervised Learning = Robust Multi-View Learning" I. Muslea, S. Minton, C. Knoblock *Proceedings19th ICML*, 2002.

[Oard 1997] "Alternative Approaches for Cross-Language Text Retrieval," D. W. Oard, in *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05. http://www.clis.umd.edu/dlrg/filter/sss/papers

[Ogawa 1995] "A new characterbased indexing organization using frequency data for Japanese documents" Y. Ogawa *Conference on Research and Development in Information Retrieval*, ACMSIGIR, Seattle, 1995, pp. 121-129.

[Pinatel, 2003] *Coloriage thématique à l'intérieur d'un document: approche contextuelle* P. Pinatel, Rapport de projet DESS RADI. Université de Caen, 2003.

[Sakamoto & *al.* 2002] "Knowledge Discovery from Semistructured Texts", H. Sakamoto, H. Arimura, & S. Arikawa, in Arikawa & Shinohara (eds) *Progress in Discovery Science* (LNAI 2281) Springer, 2002 pp. 586-599.

[Salton 1989] *Automatic text processing*, G. Salton, Addison-Wesley, 1989.

[Segond 2002] *Multilinguisme et traitement de l'information*, F. Segond (Ed.), Paris, Hermès Lavoisier, 2002.

[Stienne & Lucas 2003] "Exploitation d'information disponible sur Internet et génération d'un portail multilingue sur la thématique cinéma" N. Stienne & N. Lucas *Cide 6*, Faure et Madelaine (eds) *Document électronique dynamique*, Paris, Europia, 2003, pp. 239-255.

[Toussaint 2004] "Extraction de connaissances à partir de textes structurés" Y. Toussaint *Document numérique* Vol.8: 3, (2004) pp. 11-34.

[Valette 2004] « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet » M. Valette *CIDE.7*, P. Enjalbert, M. Gaio (eds), pp. 215-230.

[van Dijk 1988] *News analysis: case studies of international and national news in the press.* T. A. van Dijk Hillsdale, N. J., L. Erlbaum, 1988.

[Vergne 2001] Analyse syntaxique automatique de langue: du combinatoire au calculatoire. J. Vergne, *TALN 2001,* Tours vol. 1 (pp. 15-29).

[Virbel & Pascual 1996] "Semantic and Layout Properties of Text Punctuation" J. Virbel & E. Pascual, In *Proceedings of the workshop on Punctuation in Computational Linguistics*, ACL Conference, Santa Cruz, 1996.

[Yamada 1936] *Nihon bunpôgaku gairon* [*Somme sur la grammaire japonaise*] Y. Yamada, Tôkyô, Hôbunkan, 1936 (ré-imp. 1989).

ANNEXE Le vin jaune Segmentation thématique Hernandez & Grau 2002.

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.