Vers une typologie basée sur les mécanismes discursifs

Delphine Battistelli — Marie Chagnoux

LaLICC, FRE 2919, CNRS
Université Paris-Sorbonne
Maison de la Recherche
28 rue Serpente
F-75006 Paris
{delphine.battistelli, marie.chagnoux}@paris4.sorbonne.fr

Introduction

Les typologies de textes ou de corpus classiquement retenues peuvent être qualifiées soit de « typologies de surface », opposant ainsi romans, articles de presse ou articles scientifiques par exemple, soit de « typologies en profondeur », opposant descriptions, narrations ou commentaires comme le propose par exemple (Adam 92).

Dans le cadre de l'analyse de la temporalité dans un corpus hétérogène, nous avons constaté que ces typologies *a priori* n'étaient pas pertinentes[1] : en examinant les procédés linguistiques sous-jacents à la cohérence supposée des textes, il nous est en effet apparu (i) que la *typologie de surface* ne résistait pas à l'analyse puisque du point de vue temporel, certains mécanismes étaient identiques quel que soit le type de texte ; (ii) qu'en revanche la *typologie en profondeur*, si elle était peu adaptée pour caractériser un texte, pouvait être exploitée pour en caractériser les segments qui le constituent. Nous proposons ici de commenter ces deux points et d'exposer le mode de segmentation des textes qui découle de notre analyse. Cette dernière a été développée en se fondant sur le modèle des référentiels temporels tel que proposé dans (Desclés 95).

1. Typologies de textes et analyse temporelle : observations à partir d'un corpus hétérogène

1.1 Typologie de surface et analyse temporelle

En cherchant à modéliser la sémantique temporelle de textes *a priori* différents (selon une typologie classique dite de surface), nous avons constaté que de mêmes mécanismes de « ruptures temporelles » y étaient repérables et qu'il était donc fécond de ce point de vue de proposer un cadre conceptuel général permettant de rendre compte de ces mécanismes. Ces ruptures sont caractérisables en termes de modes de référenciation temporelle et peuvent par exemple relever de la modalité ou de la prise en charge (discours directs ou rapportés, médiations, commentaires, narrations, …). au sein d'un même texte. Qu'il s'agisse d'articles scientifiques, de romans, d'œuvres poétiques ou encore de pièces de théâtres, de mêmes types d'indices linguistiques sont en effet convoqués pour marquer de telles ruptures énonciatives. Ils dénotent des *mécanismes cognitifs* sous-jacents d'organisation des textes en référentiels temporels différents qu'il convient d'articuler entre eux pour rendre compte de la sémantique référentielle temporelle globale d'un texte (Desclés 95, Battistelli et al. 06). Actuellement, nous nous intéressons plus particulièrement à la description des *mécanismes linguistiques* propres au système de référenciation temporelle dans les textes en vue d'en proposer une typologie. Cette description est rendue opératoire à partir d'une distinction que nous établissons entre marqueurs temporels « cohésifs » et « incohésifs ». Les mécanismes linguistiques sont alors décrits au travers de règles faisant intervenir ces différents types de marqueurs. Des mécanismes cognitifs, modes d'organisation en référentiels temporels, et des mécanismes linguistiques, modes d'organisation de marqueurs linguistiques, participent à l'élaboration d'une typologie de *mécanismes discursifs* à l'œuvre dans l'interprétation de la temporalité de textes appartenant à des genres *a priori* différents.

1.2 Typologie en profondeur et analyse temporelle

Ainsi, dans le cadre de l'analyse de la sémantique référentielle temporelle de textes, il nous est apparu incontournable tant sur le plan théorique que sur le plan d'analyses relevant du TAL, non pas de retenir une typologie *a priori* de textes sur laquelle se fonder pour mettre en oeuvre tel ou tel type de traitement temporel, mais de proposer une typologie de mécanismes discursifs, car : (i) étant susceptibles d'intervenir dans tout type

[1] Cf. (Chagnoux 2006).

de textes ; (ii) permettant en outre d'articuler entre eux différents types de segments textuels. Cette approche rejoint celle proposée par (Smith 01), qui dégage une classification de passages discursifs et propose cinq modes (narratif, descriptif, rapport, informatif et commentaire). Ceux-ci sont eux-mêmes explicitement associés à différents types de progression temporelle mais l'articulation entre segments textuels en elle-même n'est pas décrite (une relation de succession semble implicite). Nous proposons, pour notre part, (i) d'associer non seulement aux différents types de segments repérés des modes d'organisation temporelle différents (situations en relation temporelle médiate ou immédiate avec le processus énonciatif, situées ou non dans une chronologie, …)[2], mais aussi (ii) de décrire explicitement les relations existant entre segments. Nous retenons en l'état actuel de nos analyses les relations d'*inclusion* et de *succession* entre segments textuels puisque celle de *chevauchement* n'a pas été mise en lumière jusqu'ici[3]. C'est au point (ii) que nous nous intéressons plus particulièrement ci-après.

2. Segmentation d'un texte en fonction d'indices de ruptures discursives

2.1. Le cas des ruptures de prises en charge

Pour illustrer notre propos, nous proposons d'examiner deux extraits de textes relevant de deux genres *a priori* distincts (un extrait de roman et un extrait d'article scientifique), en montrant d'une part en quoi ils illustrent des mécanismes généraux et d'autre part en quoi ils présentent des segments textuels hétérogènes.

En mars 1995, les physiciens rassemblés en hâte au Laboratoire Fermi, près de Chicago, assistèrent à un événement historique : les différentes équipes, chargées chacune d'un détecteur installé sur l'accélérateur, annoncèrent la découverte du quark top. Avec cette découverte de l'une des pièces manquantes du Modèle standard de la physique des particules, une quête de près de 20 ans s'achevait.
Le quark top est le sixième et, probablement, le dernier quark. Avec les leptons - l'électron et les particules du même type -, les quarks sont les briques fondamentales de la matière. Les quarks les plus légers, nommés u et d (pour up et down), sont les constituants des protons et des neutrons qui, en se liant à des électrons, forment tous les éléments de la Classification périodique de Mendeleïev. Les quarks plus lourds, comme les quarks c (charmé), s (strange), t (top) et b (bottom) et les leptons lourds, bien qu'abondants juste après le Big Bang, ne sont produits aujourd'hui que dans les accélérateurs de particules. Selon le Modèle standard, qui décrit les interactions entre ces briques constitutives, les leptons et les quarks se groupent en paires, souvent nommées familles.

Texte 1. Extrait de « Echos d'une crise centenaire », J.-J. Gaudillière *La Recherche*, n° 339

Le 10 septembre, je partis joyeusement pour Laubardon. Je m'embarquai à Uzerche, au petit matin, et je descendis à Bordeaux, car, avais-je écrit à Zaza, "je ne peux pas traverser la patrie de Mauriac sans m'y arrêter". Pour la première fois de ma vie, je me promenais seule, dans une ville inconnue. Il y avait un grand fleuve, des quais brumeux, et déjà les platanes sentaient l'automne. Dans les rues étroites, l'ombre jouait avec la lumière ; et puis de larges avenues filaient vers des esplanades. .

Texte 2. Extrait des *Mémoires d'une jeune fille rangée*, S. De Beauvoir[4]

Il est possible de repérer dans chacun des extraits une rupture qui articule deux segments temporellement homogènes, c'est-à-dire des segments où les relations temporelles entre propositions relèvent d'un certain type d'organisation propre au référentiel auquel elles appartiennent : un premier segment qui s'apparente à de la narration (où alternent passés simples et imparfaits) et un second segment qui peut être décrit comme définitoire (au présent, dans le premier extrait) ou comme descriptif (à l'imparfait, dans le second extrait). Un certain nombre d'indices permettent de justifier ces ruptures : les indices calendaires, comme « en mars 1995 » ou « le 10 septembre », ne sont par exemple présents que dans la partie narrative ; la partie définitoire du premier extrait est quant à elle constituée d'une succession de verbes statifs ; *etc*.

Du point de vue de la prise en charge énonciative, les deux textes proposent deux mécanismes de rupture différents : dans le premier extrait, la rupture est introduite par « Selon » alors que, dans le second, elle est signalée par la présence de guillemets. Nous ne développons pas ici ce que chacune de ces citations a de

[2] A chaque référentiel sont associés un certain nombre de propriétés qui interviennent dans l'interprétation. On distingue principalement : (i) le référentiel énonciatif sur lequel les situations s'organisent par rapport au processus énonciatif ; (ii) le référentiel non-actualisé sur lequel les situations sont en rupture avec ce processus ; (iii) le référentiel du possible sur lequel les situations peuvent être contrefactuelles ou réelles ; (iv) le référentiel des vérités générales sur lequel les situations sont vraies à tout instant, etc.

[3] Les relations entre segments exhibées ici ne sont donc pas des « relations discursives » au sens de celui développé dans le cadre de la SDRT (Asher et al. 93) par exemple.

[4] Gallimard Folio, ed. 1987, p. 351.

spécifique – appel à une norme dans le premier extrait et auto-citation dans le second – pour ne retenir que le principe de rupture de prise en charge marquée dans les textes.

Le fait que ces mécanismes d'interprétation sont mis en place au regard de l'opposition entre marqueurs temporels « cohésifs » et « incohésifs » permet d'envisager un traitement automatique de reconnaissance de ces segments.

2.2 Marqueurs temporels cohésifs et incohésifs

Dans l'analyse temporelle de surface d'un texte, il peut en effet être opportun de distinguer deux types de marqueurs linguistiques : des marqueurs de « cohésion » et des marqueurs d' « incohésion ».

Les marqueurs de cohésion assurent l'homogénéité d'un point de vue temporel d'un segment de texte en inscrivant les situations :

- *Cas 1* : soit dans une chronologie (ex. des narrations) : c'est par exemple le cas des connecteurs, des adverbes, des temps verbaux concordant en systèmes (système d'alternance de verbes à l'imparfait et au passé simple par exemple), *etc* ;
- *Cas 2* : soit dans une « vision statique » (ex. des commentaires ou des définitions) : comme c'est le cas par exemple pour une suite de verbes statifs conjugués au présent ou à l'imparfait, et où, bien que possible, le calcul des relations temporelles n'est pas pertinent[5].

Les marqueurs d' « incohésion » indiquent les ruptures. Ils peuvent être soit de simples marqueurs comme les indices typographiques tels que les guillemets, suivis ou non de deux points, soit des ensembles de marqueurs, soit encore des temps ou des modes verbaux. Ces ruptures peuvent être[6] :

- liées à un changement référentiel temporel (énonciatif, possible, modal…) ;
- liées à une « violation » du principe d'homogénéité des temps ou modes verbaux : dès lors, tout en étant éventuellement pris en charge par le même énonciateur, des segments textuels peuvent être identifiés comme ne s'inscrivant plus dans la continuité référentielle[7] des situations décrites précédemment dans le texte (c'est ainsi le cas quand à une suite de temps du passé succède une suite de présents).

Cette distinction renvoie donc à une typologie d'un certain type de « ruptures discursives » à l'œuvre dans les textes. Les mécanismes discursifs qui sont sous-jacents à cette typologie sont décrits au travers de règles systématiques qui conduisent à découper les textes en *segments textuels* ; ces derniers se trouveront en relation d'inclusion ou de succession au regard des types d'indices cohésifs ou incohésifs en jeu.

2.3 Relations entre segments textuels

Ainsi, par exemple, quand un marqueur d'incohésion lié à un changement de prise en charge énonciative est repéré, un début de segment *Seg_i* est signifié et n'est refermé qu'au moment où un autre marqueur d'incohésion énonciative est repéré ou quand des marques typographiques telles que par exemple les guillemets de fin apparaissent ; c'est donc la relation de succession entre segments textuels qui prévaut ici, sauf dans deux cas : (i) quand des énonciateurs semblables sont identifiés[8] ; (ii) dans le cas de l'énonciateur principal : étant donné qu'un énonciateur principal est toujours présent, que ce soit de manière explicite ou non, le texte entier correspondra lui-même à un segment et inclura donc tous les autres forcément. Il n'en est pas de même dans le cas des marqueurs d'incohésion liés à une violation du principe d'homogénéité des temps ou modes verbaux : c'est la relation d'inclusion qui prévaut alors, au sens où si un segment temporel cohésif *Seg_i* est ouvert, il le restera par défaut même si un autre segment temporel cohésif *Seg_i+1* est identifié ; ce dernier se trouvera alors dans une

---

[5] En d'autres termes, certains marqueurs invitent nécessairement le lecteur à calculer des relations temporelles entre les procès pour comprendre le texte, d'autres non. Les cas 1 et 2 renvoient respectivement dans la terminologie classique au type 'narration' et au type 'commentaire'.

[6] Ici, nous nous intéresserons uniquement aux mécanismes propres au référentiel énonciatif sachant que, dans le cadre d'applications visant à identifier « qui », « quand » et « où », il est particulièrement pertinent de pouvoir traiter ce type de référentiel ; à un même texte peuvent en effet être associés différents référentiels énonciatifs qui renvoient aux différents énonciateurs co-présents dans le texte (énonciateurs premiers, énonciateurs seconds, …).

[7] A propos de l'enjeu que constitue le problème d'assurer une continuité référentielle entre segments textuels extraits (et des solutions envisagées pour tenter d'y répondre) pour les systèmes de résumé automatique en particulier, voir par exemple (Battistelli et Minel 06).

[8] L'application actuelle que nous avons développée ne permet pas de réaliser cette tâche. Notre objectif s'est tout d'abord limité uniquement au repérage des ruptures discursives et non à l'identification automatique des entités nommés pour répondre à la question "qui a dit" comme proposé par (Daille *et* Morin 00).

relation d'inclusion avec le précédent. L'extrait 3, tiré d'un article de presse, permet d'illustrer le cas de plusieurs ruptures énonciatives ; trois segments, *seg_1*, *seg_2* et *seg_3* et donc trois énonciateurs différents, peuvent être dégagés en se basant sur les indices d'incohésion, marqués en gras dans le texte.

<seg_1>A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens **affirment qu'<seg_2>** ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise. **</seg_2> De son côté**, la police libanaise **soutient que <seg_3>**la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même**</seg_3></seg_1>**

Texte3. Extrait de "L'armée israélienne investit le Liban sud."[9]

L'énonciateur principal prend en charge le contenu de *seg_1 :* « *A propos du lieu de l'enlèvement des deux soldats, les versions diffèrent. Les Israéliens affirment qu'[…]. De son côté, la police libanaise soutient que […]* » ; la prise en charge de *seg_2 :* « *ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière libanaise* » est attribuée à « *Les Israéliens* » ; la prise en charge de *seg_3* : « *la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la frontière libano-israélienne où une unité israélienne avait pénétré le matin même*» est attribuée à « *la police libanaise* ».

Le repérage automatique des segments est réalisé grâce à une base de marqueurs couplée à une base de règles au sein de *Chronotexte*[10], le système ne permet pas pour l'instant de prendre en compte la structure hiérarchisée des segments et indique seulement les ruptures entre segments. Nous travaillons actuellement au balisage de la structure selon la DTD définie par (Couto 06). Ainsi le texte 3 est balisé par le fichier présenté dans la figure 1.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE Texte SYSTEM "DocumentNaviTexte.dtd">
<Texte>
  <Tete>
    <Sequence Type="Regroupement" Nro="1" >
      <UTP Type=" Referentiel" Nro="1"/>
      <UTP Type=" Referentiel" Nro="3"/>
    </Sequence></Tete>
  <Corps>
    <UT Type="Referentiel" Nro="1">
        <Attribut Nom="Nature">Enonciatif_global</Attribut>
        <Attribut Nom="Enonciateur">Wikipedia</Attribut>
    <UT Type=" Proposition" Nro="1">
        <Chaine>A propos de l'enlèvement des deux soldats, les versions diffèrent</Chaine></UT>
      <UT Type=" Proposition " Nro="2">
        <Chaine>Les Israéliens affirment qu'</Chaine></UT></UT>
      <UT Type=" Référentiel " Nro="2">
        <Attribut Nom="Nature">Enonciatif_local</Attribut>
        <Attribut Nom=" Enonciateur ">Les Israéliens</Attribut>
      <UT Type=" Proposition " Nro="3">
        <Chaine>ils ont été capturés près de la ferme collective de Zarit en territoire israélien tout près de la frontière
libanaise.</Chaine></UT></UT>
      <UT Type=" Référentiel " Nro="3">
        <Attribut Nom="Nature">Enonciatif_glocal</Attribut>
        <Attribut Nom=" Enonciateur ">Wikipedia</Attribut>
      <UT Type=" Proposition " Nro="4">
        <Chaine>De son côté, la police libanaise soutient que</Chaine></UT></UT>
      <UT Type=" Référentiel " Nro="4">
        <Attribut Nom="Nature">Enonciatif_local</Attribut>
        <Attribut Nom=" Enonciateur ">la police libanaise</Attribut>
      <UT Type=" Proposition " Nro="5">
        <Chaine>la capture s'est produite dans la région de Aïta al-Chaab en territoire libanais donc, proche de la
frontière libano-israélienne></Chaine></UT>
      <UT Type=" Proposition " Nro="6">
        <Chaine>où une unité israélienne avait pénétré le matin même.></Chaine></UT></UT>
  </Corps> </Texte>
```

Figure 1. *Fichier XML associé à l'extrait de texte 3*

[9] Arsenault C., http://www.rfi.fr/actufr/articles/079/article_45021.asp, (publié le 12/07/06 et consulté le 15/09/06).
[10] *Chronotexte* est présenté en détail dans (Chagnoux 06).

Le *Corps* contient les différentes propositions du texte. Chaque segment est défini grâce à une *Unité Textuelle* (UT) à laquelle est associé un attribut *Nature* qui définit le type de référentiel (ici, énonciatif global ou énonciatif local) et, dans le cas d'un référentiel énonciatif, un attribut *Enonciateur*. Il faut noter que chaque type de référentiel convoque des attributs différents permettant d'encoder toutes les propriétés qui lui sont associées. La *Tete* permet de traiter le problème de la discontinuité : l'opération de *Séquence* rétablit la cohésion entre des segments discontinus dans le texte. Ainsi, *seg_1* apparaît comme un segment qui inclut *seg_2 et seg_3*.

Conclusion

Le balisage des différentes ruptures discursives repérées (énonciatives, modales,…) permet de proposer une autre représentation du texte. Intégrée à un outil de navigation textuelle par exemple[11], elle peut ainsi faire explicitement apparaître des mécanismes discursifs qui relèvent d'une « typologie de (segments) de textes » et donc d'une « typologie infra-textuelle ». Cette représentation peut alors à son tour être utilisée en vue de distinguer les textes :

(i) selon la complexité avec laquelle ils comptent tels ou tels types de segments (nombre de commentaires, de définitions, de médiations, …) ou encore d'énonciateurs ;

(ii) mais aussi selon la complexité avec laquelle ils articulent des segments homogènes (taux de relations de succession ou d'inclusion entre citations, taux de relations de succession ou d'inclusion entre narrations et commentaires, …).

Ce type d'évaluation quantitative de mécanismes discursifs (rendue seulement possible par des outils de TAL) peut alors être vu comme fournissant les bases d'une éventuelle autre typologie de textes, participant elle aussi selon nous à problématiser la notion théorique de « texte ».

Bibliographie

Adam J.-M, *Les Textes Types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris: 1992. Université / Seuil.

Asher N, Busquets J., Vieu L., « La SDRT: une approche de la cohérence du discours dans la tradition de la sémantique dynamique », *Verbum* 23, 73-101, 1993.

Battistelli D., Chagnoux M., Desclés J.-P., « Référentiels et ordonnancements temporels dans les textes », *Cahiers Chronos*, 2006.

Battistelli D., Minel J.-L., « Les systèmes de résumé automatique : comment assurer une continuité référentielle dans la lecture des textes », in G. Sabah (Ed.), *Compréhension des langues et interaction*, p. 295-330, 2006.

Chagnoux M., *Temporalité et aspectualité dans les textes français : modélisation sémantico-cognitive et traitement informatique*, Thèse de doctorat de l'Université Paris IV – Sorbonne, 2006.

Chagnoux M., « La cohérence temporelle dans les textes », *Actes Regards Croisés sur l'Unité Texte*, 17-20 avril 2004, Chypre.

Couto J., *Modélisation des connaissances pour une navigation textuelle assistée. La plate-forme logicielle NaviTexte*. Thèse de doctorat, Université Paris-Sorbonne, 2006.

Couto J., Minel J.-L, « SEXTANT, un langage de modélisation des connaissances pour la navigation textuel», *ISDD'06*, Caen, p. 80-90, 2006.

Daille B., Morin E., « Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations », *Traitement Automatique des Langues (TAL)*, 41(3):601-622, 2000.

Desclés J.-P., « Les référentiels temporels pour le temps linguistique », *Modèles linguistiques*, XVI (2), 9-36, 1995.

Smith C., « Discourse modes: aspectual entities and tense interpretation », *Cahiers de grammaire*, 26, 183-206, 2001.

Wonsever D., *Repérage automatique des propositions par exploration contextuelle*, Thèse de doctorat, Université Paris-Sorbonne, Paris, 2004.

[11] Nous travaillons actuellement à cette intégration dans le cadre de la plate-forme *NaviTexte* (Couto 06, Couto et Minel 06). La visualisation de la dynamique référentielle du texte se présente sous la forme d'un graphe orienté dont les nœuds sont étiquetés d'une information sur le référentiel concerné et sur l'ensemble des propositions du texte qui y sont associées. Il est à noter que notre démarche présuppose un découpage en propositions du texte. Nous utilisons pour cette tâche le système développé par (Wonsever 04).