Améliorer le découpage en phrases sous INTEX

Nathalie FRIBURGER, Anne DISTER, Denis MAUREL

Abstract : Dans cet article, nous proposons de voir comment il est possible d'améliorer le découpage d'un texte en phrases sous INTEX. Dans un premier temps, nous aborderons les problèmes posés par le point, signe ambigu qui peut marquer la fin d'une phrase ou non. Ensuite, nous décrirons les possibilités et nos choix de découpage pour le reste de la ponctuation. Nous présenterons pour terminer les résultats obtenus sur un triple corpus après les améliorations apportées par nos graphes.

Keywords : tokenization, sentence, INTEX, natural language processing

Mots clés : découpage, phrase, ponctuation, INTEX, traitement automatique des langues

1. Le découpage automatique de phrases

À la lecture d'un texte, nous réalisons inconsciemment un certain nombre de tâches qui vont nous permettre de le comprendre. Aidés par la ponctuation et le sens, nous découpons le texte en phrases. Cette tâche constitue une première phase de traitement que l'on peut automatiser grâce au logiciel INTEX. Elle est déterminante pour la suite du traitement du texte qui se basera toujours sur ce découpage préliminaire en unités élémentaires.

Ce découpage est difficile, mais les automates à états finis d'INTEX permettent de décrire ce qui caractérise les fins de phrases et

✉ Nathalie FRIBURGER, Denis MAUREL :
LI (Laboratoire d'Informatique de Tours) friburger@univ-tours.fr
maurel@univ-tours.fr
Anne DISTER : Université de Liège, Quai Roosevelt 1b, B-4000 Liège, Belgique
Fax : 0032 4 366 57 84 adister@ulg.ac.be

ce qui ne les caractérise pas. L'automate de découpage des phrases, nommé *sentence*, insère dans le texte une étiquette {S} qui symbolise la séparation entre deux phrases.

Un tel automate ne permet pas de délimiter toutes les phrases comme le ferait le lecteur humain ; il y a quelques limites inhérentes à l'automatisation de cette tâche. Des choix personnels se posent aussi.

Nous avons constaté que le découpage en phrases qui se fait actuellement sous INTEX pouvait être amélioré et nous proposons dans cet article de présenter les modifications que nous avons apportées au graphe *sentence* [1].

Lorsqu'on automatise le découpage en phrases, on peut craindre deux types d'erreurs : les bruits (trop d'étiquettes {S}, et donc trop de phrases) et les silences (des étiquettes manquantes, et donc, des phrases non repérées). Dans notre travail, la ligne de conduite générale a été d'éviter le bruit. En effet, dans les phases de traitement ultérieures, il est plus gênant sous INTEX d'avoir des phrases en surnombre que l'inverse. C'est le cas notamment pour la localisation automatique de motifs, qui ne peut se faire qu'à l'intérieur d'une phrase, entre deux symboles de fin de phrase.

D'une manière générale, nous avons amélioré le découpage en élargissant les contextes gauche et droit des motifs. Nous utilisons des étiquettes morphologiques : *<MAJ>*, *<MIN>*, *<PRE>*, *<NB>* [2], etc. ; nous prenons en compte des ponctuations : *. ; , ! ?*, etc. ; nous utilisons également le lexique. De plus, la possibilité de formaliser l'absence d'un blanc grâce au symbole # sera très utile pour décrire des motifs comme les sigles.

Voici notre graphe sentence :

---

[1] Cet article s'adresse plus particulièrement aux utilisateurs d'INTEX afin de leur permettre de comprendre les erreurs résiduelles de découpage qu'ils pourraient rencontrer dans un texte en utilisant notre graphe — et éventuellement, comment y remédier. Pour une description plus complète des procédures de découpage, voir DISTER 1997.

[2] <PNC> correspond à tout symbole de ponctuation, <MAJ> est un mot en majuscule, <PRE> est un mot en minuscule commençant par une majuscule, <MIX> est un mot comprenant des majuscules et des minuscules mélangées (ex : *McDonald*) et <NB> est un nombre (écrit en chiffres).

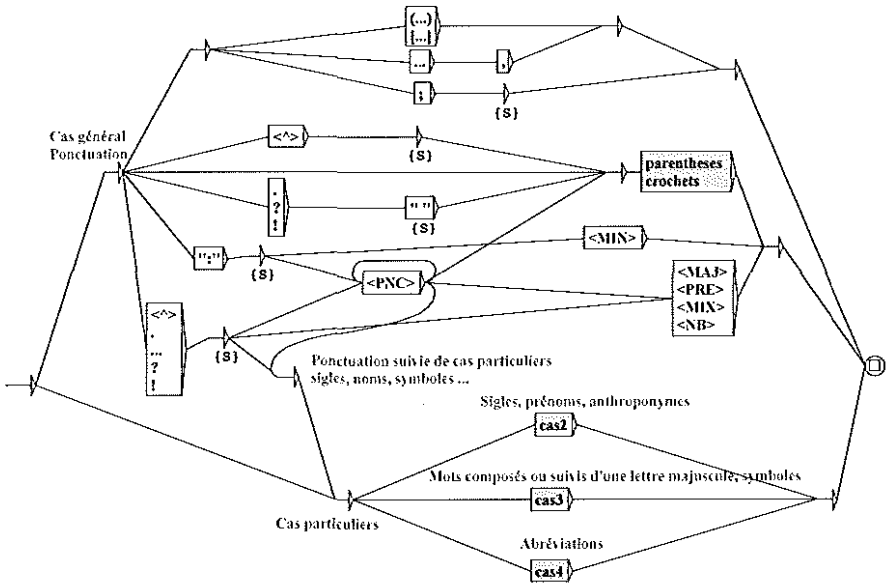

Figure 1. *Le graphe sentence*

2. Ambiguïté du point

Le découpage des phrases lorsqu'il n'est pas ambigu est décrit dans la partie supérieure du graphe *Sentence*. Si on découpe les phrases en n'utilisant que cette partie du graphe (pas de traitement des ambiguïtés), on obtient un très mauvais découpage.

J'ai rencontré M. {S}Dupont à la poste.
Pour le syndicat de la F.{S}E.{S}N.{S}, c'est une erreur.

Le point est un signe ambigu en présence de majuscules. Cette question a déjà été abordée par (SILBERZTEIN 1993) et (DISTER 1997) ; les motifs contenant des majuscules et des points sont de quatre types :

- Motifs anthroponymiques : précédés de l'abréviation de *Monsieur*, *M.* ou *MM.* au pluriel (*M. Dupont* ou *MM. Dupont et Dupond*), contenant un prénom abrégé (*J. Dupont*) ou pour ces deux raisons à la fois (*M. J. Dupont*)
- Sigles, lorsque l'ancienne notation est utilisée :

La S.N.C.F. gère les chemins de fer français.

Notons que la nouvelle notation est sans points : *La SNCF gère les chemins de fer français.*

- Symboles et mots composés avec une majuscule

  Ce timbre coûte 20 F. Il a été acheté chez un philatéliste.
  Ces aliments contiennent de la vitamine A, B et C. Durant me l'a confirmé.

- Abréviations diverses

  éd. Gallimard
  Chap. 4
  Cf. France-Italie en juin 2000.

Ces différents cas sont traités respectivement dans les sous-graphes *cas2* (pour les motifs anthroponymiques et sigles), *cas3* (pour les symboles et mots composés avec une majuscule) *et cas4* (pour les abréviations) du graphe *sentence* (Fig. 1). Comme dans tous les graphes d'INTEX, les cases grisées symbolisent l'appel à un sous-automate.

2.1. Le point, les sigles et les anthroponymes

Nous traitons ensemble les sigles et les anthroponymes. En effet, un sigle de deux lettres se trouvant en fin de phrase est ambigu avec un anthroponyme comme dans :

Ils ont innocenté O.J. Simpson du meurtre de sa femme.

Le graphe *cas2* présente l'automate qui évite l'insertion d'un séparateur de phrases dans les sigles et les motifs anthroponymiques.

Le sous-graphe *MotifAnthro* (Fig. 3) décrit les motifs anthroponymiques non ambigus avec des sigles. Le point qui suit *M*, *MM* ou *Prof* n'est pas reconnu comme un point final s'il est suivi d'un prénom abrégé ou d'un mot commençant par une majuscule. *PhTh* est le nom du sous-graphe qui liste les prénoms abrégés à partir de la première voyelle comme *Th* pour *Thierry*, *Ph* pour *Philippe*, etc. et le sous-graphe *Prénoms* décrit les différents motifs possibles pour un prénom.

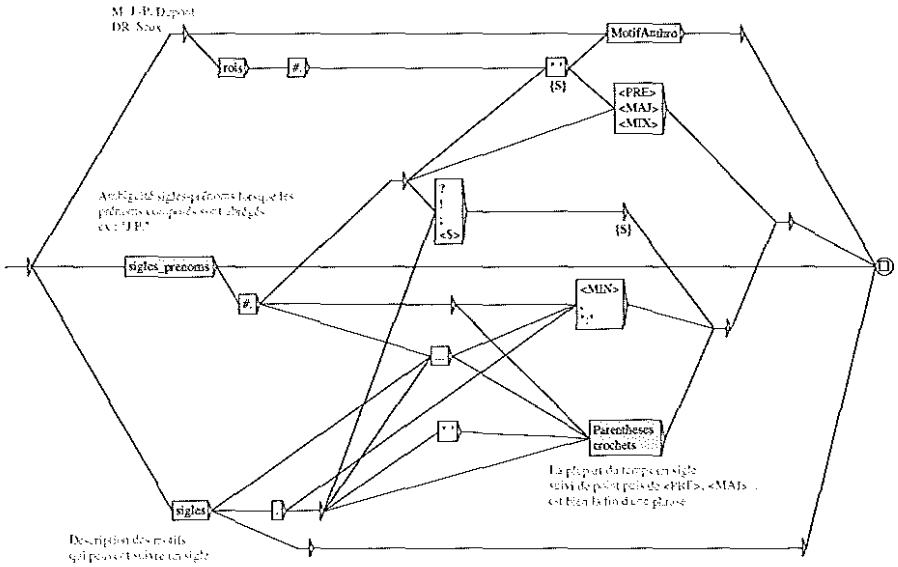

Figure 2. *Le graphe cas2*

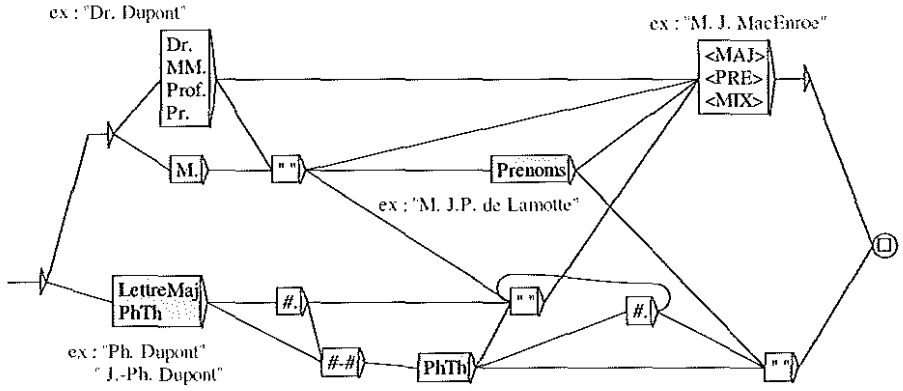

Figure 3. *Le graphe MotifAnthro*

Le sous-graphe *rois* du graphe *cas2* liste des prénoms de rois qui, suivis des chiffres romains *V* ou *X* puis d'un point final, sont ambigus avec un prénom de personne abrégé en *X* (pour *Xavier*, *Xénophon*, etc.) ou en *V* (pour *Valère*, *Victor*, etc.). Il s'agit, dans la phrase qui

suit, de repérer que l'on parle de Charles X — en fin de phrase, on insère donc la marque {S} — et non de Xavier Dupont (*X. Dupont*), auquel cas il ne faut pas insérer de marque de fin de phrase.

Cela s'est passé à la cour de Charles X. {S} Dupont ne le savait pas.

Les sigles sont eux aussi traités dans le graphe *cas2*. Les sigles de deux lettres ambigus avec un prénom abrégé composé[3] sont traités dans le graphe *sigle_prenoms*. Tous les autres sigles (*C.F.D.T., S.N.C.B.*, etc.) sont reconnus par le graphe *sigles* (Fig. 4) dans lequel nous avons précisé qu'entre deux lettres composant un sigle il ne devait pas y avoir d'espaces (le symbole # permet de formaliser l'absence de blancs sous INTEX). On remarquera que le dernier point d'un sigle n'est pas pris en compte dans ce graphe puisque c'est celui-ci qui peut être ambigu avec le séparateur de phrase comme dans l'exemple ci-dessous :

La C.F.D.T. a lancé une grève.
La grève a été lancée par la C.F.D.T.

L'automate *sigles* décrit aussi les compositions de sigles telles que *C.G.T.-C.F.D.T.*

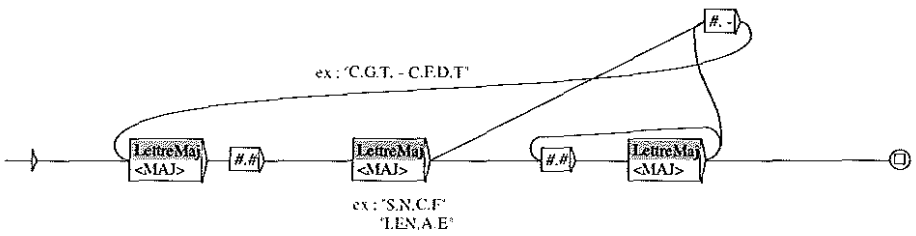

Figure 4. *Le graphe sigles*

Nous éviterons aussi de placer une fin de phrase entre un sigle et une information entre parenthèses qui le suit comme dans :

[3] Ceci est particulièrement vrai pour les prénoms anglo-saxons : *O.J. Simpson, F.D. Roosevelt,* etc. Par ailleurs, il n'est pas rare de rencontrer dans les textes des abréviations de prénoms composés français dans lesquelles on omet le tiret. On a ainsi *J.C.* pour *Jean-Charles*, *P.Y.* pour *Pierre-Yves*, etc.

S.N.C.F. (Société Nationale des Chemins de fer Français).

2.2. Le point et les symboles

Les symboles composés d'une seule lettre majuscule, les abréviations d'unités de mesure et les symboles monétaires (*V* : *volt*, *F* : *Franc*, etc.) ne posent pas de problème à l'intérieur d'une phrase. En effet, dans ce cas, ils ne sont pas suivis d'un point (au contraire des sigles).

Un Magritte de 14 000 F a été volé au Centre Pompidou.
Il faut une ampoule de 100 W pour cette lampe.

Mais lorsqu'ils sont en fin de phrase, les symboles sont suivis d'un point qui rend la séquence ambiguë : le symbole abrégé en une lettre majuscule en fin de phrase suivi d'un point pourrait être analysé comme l'initiale d'un prénom, et le mot en majuscules qui commence la phrase suivante comme un nom de famille. Suivant notre graphe de reconnaissance des motifs anthroponymiques, aucune étiquette de fin de phrase ne serait insérée entre l'initiale et le nom erronément reconnus comme tels :

C'est un Magritte de 14 000 F. Volé au Centre Pompidou, il ne sera sans doute jamais retrouvé.

Pour pallier ce type de problèmes, nous avons recensé dans le graphe cas3 (Fig. 5) tous les cas où une lettre majuscule provoque une ambiguïté lorsqu'elle apparaît en fin de phrase.

Le sous-graphe *MotifSymboles* (Fig. 6) reprend les motifs que l'on peut trouver autour d'un symbole. Nous y décrivons l'emploi du symbole monétaire *F* et des autres symboles (dans le sous-graphe *Symboles1Maj*) ; c'est la présence d'un nombre avant le symbole qui va permettre de désambiguïser ce cas avec le cas d'un motif anthroponymique, et d'insérer correctement le symbole de fin de phrase.

C'est un Magritte de 14 000 F. {S} Volé au Centre Pompidou, il ne sera sans doute jamais retrouvé.

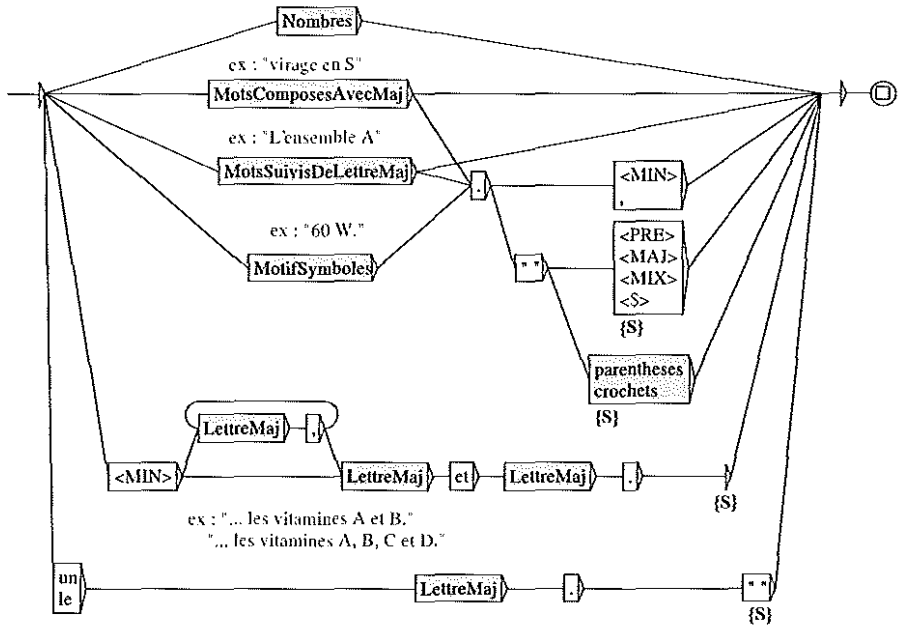

Figure 5. *Le graphe cas3*

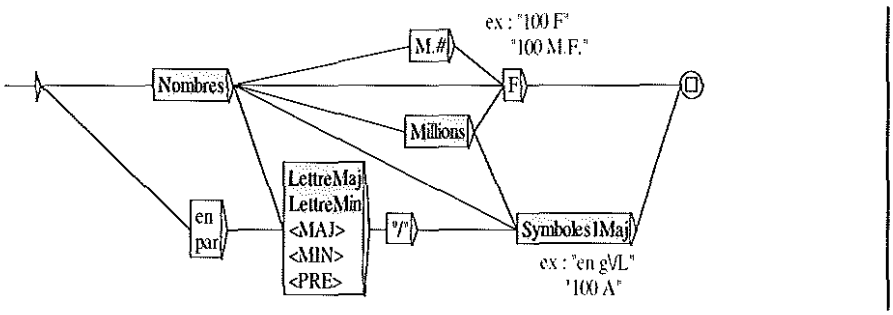

Figure 6. *Le graphe MotifSymboles*

Le sous-graphe *MotsComposésAvecMaj* rassemble une liste de mots composés qui contiennent une majuscule (*linéaire B*, *film X*, *série Z*, etc.) tandis que le sous-graphe *MotsSuivisDeLettreMaj* contient des listes de mots qui ne sont pas des mots composés mais

qui peuvent être suivis d'une majuscule (*variable*, *ensemble*, *équipe*, *pool*, *bac*, etc.).

Il se peut aussi que l'on rencontre une unité telle que g/L (concentration en gramme par litre) ; ce motif composé d'une lettre, de la barre oblique /, d'une lettre majuscule et d'un point est lui aussi reconnu.

La concentration en sel est de 110 g/L. {S} Les résultat sont étonnants !

Des séquences telles que

vitamines A, B et C.
un A.
le B.

sont également prises en compte.

2.3. Le point et les abréviations

Le point abréviatif est ambigu puisqu'en fin de phrase il absorbe le point final : la fin de phrase n'est donc pas formellement marquée. On distinguera deux types d'abréviations : celles qui ne peuvent jamais se trouver à la fin d'une phrase et dont le point ne peut être ambigu avec le point final, et celles qui peuvent se trouver à la fin d'une phrase. Des abréviations comme *cf.*, *i.e.* ou *c.-à.-d.* sont nécessairement suivies d'une précision ou d'une indication et ne peuvent être la fin d'une phrase. Par contre des abréviations telles que *etc.*, *ap. J.-C.*, peuvent être ou ne pas être en fin de phrase. Une abréviation qui peut se trouver en fin de phrase n'en est pas une si elle est suivie d'un mot en minuscule ou d'une ponctuation qui ne termine pas la phrase telle que la virgule, le double point, la parenthèse ouvrante ou fermante. Par contre, si le mot qui suit est un nombre ou s'il commence par une majuscule, on ne peut choisir automatiquement si l'on est en fin de phrase. Nous avons préféré éviter de mettre une marque de fin de phrase lorsque cela risque de créer trop fréquemment du bruit. Dans des cas comme ceux que l'on vient de citer, seul le jugement humain ou une description lexicale trop complète pour être possible peut décider. C'est pourquoi nous avons aussi tenu compte des abréviations qui doivent ou peuvent être suivies d'un nombre afin d'empêcher une fin de phrase avant ce nombre.

Le recensement de toutes les abréviations graphiques (reprises dans *cas4*) pose problème car on ne peut en faire une description exhaustive. Beaucoup sont par ailleurs non codifiées et relèvent de la création des scripteurs. Nous n'avons recensé que les plus fréquentes de la langue générale.

3. La ponctuation : choix de découpage

Les signes de ponctuation non terminaux sont les suivants : la virgule, les deux points, les guillemets, les parenthèses et les crochets. Les signes de ponctuation terminaux (sauf dans les cas ambigus décrits précédemment) sont les suivants : le point, le point d'interrogation, le point d'exclamation, le point virgule et le saut de paragraphe.

3.1. Parenthèses et crochets

Nous n'insérons pas de signe de séparation de phrases dans les parenthèses et les crochets afin de ne pas couper la phrase dans laquelle ces parenthèses ou ces crochets apparaissent (Fig. 7) :

> *Pierre **(vous savez comment il est : gentil, intelligent, etc.)** est l'homme de ma vie.*
> *La réaction de panique rétrospective qu'a déterminée la crise de 68, révolution symbolique qui a secoué tous les petits porteurs de capital culturel, a créé (avec, en renfort, l'effondrement — inespéré ! — des régimes de type soviétique) les conditions favorables à la restauration culturelle aux termes de laquelle la « pensée Sciences Po » a remplacé la « pensée Mao ». (Le Monde)*
> *Se conformant aux règles qui régissent les rapports entre maisons de disques, Deutsche Grammophon a pris les devants en demandant aux différents éditeurs concernés l'autorisation de faire porter ses couleurs à des chefs d'orchestre (Schuricht, Klemperer, artistes EMI et Decca ; Walter, artiste CBS, Kleiber, artiste Decca) présents dans ce coffret commémoratif d'un intérêt exceptionnel. (Le Monde)*

Dans les trois exemples précédents, on a empêché (à l'aide du graphe *parentheses*) l'insertion du symbole de fin de phrase après un signe de ponctuation terminal (deux points, point d'exclamation, etc.).

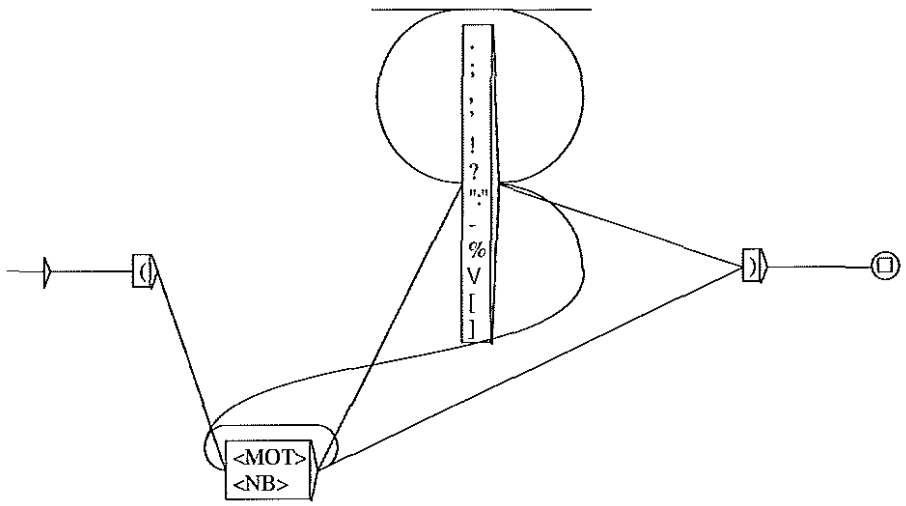

Figure 7. *Le graphe Parentheses*

Le revers de ce choix est que l'on rencontre parfois, entre parenthèses ou crochets, d'assez longs textes contenant de nombreuses phrases. On pourrait donc encore améliorer notre graphe en décidant de couper à l'intérieur des parenthèses lorsque celles-ci sont directement précédées ou suivies d'un point, comme dans l'exemple suivant :

> *Il y a ballottage.{S} [Six candidats postulaient à la succession de Jean Duroisel (RPR), récemment décédé. L'absence d'un candidat du Front national, et la présence d'un candidat unique pour la majorité départementale RPR-UDF, M. Guy Verin, ont rendu les choix relativement clairs à droite. En 1985, Jean Duroisel avait été réélu au second tour avec 2 341 voix (59,44 %) contre 1 597 (40,55 %) à Mme Dumant sur 3 938 suffrages exprimés, 4 093 votants (soit 21,55 % d'abstention) et 5 218 inscrits. Au premier tour, les résultats avaient été les suivants : inscr., 5 219; vot., 3 982; abst., 23,70 %; suffr. expr., 3 813; Mme Dumant, 1 100 voix (28,84 %); Jean Duroisel, 1 070 (28,06 %); M. Verrier, div. d., 1 027 (26,93 %); Mme Pinchon, FN, 186 (4,87 %); MM. Basquin, PC, 152 (3,98 %); Rasson, PCI, 150 (3,93 %); Kuc, div. d., 128 (3,35 %).] (Le Monde)*

3.2. Les tirets

Actuellement, notre graphe *Sentence* ne prend pas en considération les tirets. Ceux-ci peuvent être utilisés comme des parenthèses et contenir des signes de ponctuation terminaux. Ceci génère le même problème qu'au point précédent. Notons néanmoins que l'on retrouvera devant chaque tiret marquant une énumération un signe de fin de phrase, à cause du retour chariot.

3.3. Les points de suspension

On trouve les points de suspension lorsqu'on ne veut pas allonger une énumération, dans une phrase volontairement inachevée ou lorsqu'on laisse un temps avant de terminer la phrase.

Si le point de suspension est suivi d'un mot en minuscule ou d'une virgule, ce point de suspension n'est pas la fin d'une phrase.

> *Tout a commencé lorsque l'ANPE lui a proposé une place au Normandy... pour arroser les fleurs. (Ouest France)*
> *De Mel Gibson, de passage à Paris, à Roger Hanin, en passant par Alain Delon, Bernard Kouchner, le Professeur Schwartzenberg, Enrico Macias, Thierry Lhermitte..., les personnalités se pressent nombreuses au journal de la petite chaîne qui meurt. (Le Monde)*

Un problème apparaît lorsque le mot qui suit le point de suspension commence par une majuscule. Si tel est le cas, soit ce mot est un nom propre qui appartient à la phrase en cours, soit il est au début d'une nouvelle phrase, qu'il s'agisse d'un nom propre (comme c'est le cas dans notre second exemple) ou non.

> *Un colloque organisé par l'Université d'Angers à la fin de l'année dernière a bien montré l'étonnante richesse et diversité de cette littérature vendéenne souvent signée des plus grands noms : Hugo, Balzac, Barbey d'Aureyvilly, Chateaubriand, Nerval et même... Jules Verne ! (Ouest France)*
> *Après un premier acte sur des chapeaux de roue, fait surtout de répliques pas légères mais drôles, la pièce s'étire, déraille...Régis Santon n'arrange rien en laissant Marc de Jonge jouer un général de convention pure, un braillard machinal. (Le Monde)*

Les séquences « ...<MAJ> » ou « ...<PRE> » ne sont pas rares dans les textes, et nous avons donc pu observer que les points de sus-

pension étaient dans leur très grande majorité[4] à la fin de la phrase. C'est pourquoi, contrairement à notre démarche qui préfère le silence au bruit, nous avons choisi de placer une fin de phrase après le point de suspension s'il est suivi de <MAJ> ou de <PRE>[5].

3.4. Les guillemets

Sous INTEX, le symbole guillemet est le guillemet anglais " ". Ce symbole est le même, qu'il soit ouvert ou fermé. On ne peut donc pas empêcher que la phrase dans laquelle on a une citation entre guillemets soit scindée car on ne peut distinguer le cas où l'on a une citation dans une citation de celui où l'on a deux citations consécutives très proches.

> *Cet homme politique a dit : "Je pense à ce que l'on a appelé le "retour de l'individualisme", sorte de prophétie auto-réalisante qui tend à détruire les fondements philosophiques la notion de responsabilité collective."(Le Monde)*
> *Ce qui ne l'empêche pas de bénéficier du soutien des États-Unis, qui le présentent comme "un homme politique modéré" et "un ancien communiste". (Le Monde)*

La seule possibilité qui s'offre à nous est donc de placer les fins de phrases sans tenir compte de la présence des guillemets.

3.5. Les deux points

Nous avons choisi de placer une fin de phrase après les deux points lorsqu'ils sont suivis de guillemets comme dans :

> *Un intérêt qu'un aveugle d'Ecommoy, dans la Sarthe, fidèle abonné de la bibliothèque segréenne a exprimé en ces termes à l'assemblée générale de l'association :{S} " Ça me permet de passer des journées agréables.{S} Des journées qui seraient bien tristes sans cela ".(Ouest France)*

Dans tous les autres cas, nous ne plaçons pas de marque de fin de phrase, c'est-à-dire lorsque :

– on cite les éléments d'un ensemble :

---

[4] Nous n'avons pas fait de calcul exact, mais il apparaît dans notre corpus que cela arrive dans plus de 90 % des cas.

[5] Le premier des deux exemples ici cités sera donc erronément analysé et une marque de fin de phrase insérée avant *Jules Verne*.

Les trois autres quarts ont été payés par d'autres contributeurs : Arabie saoudite, Allemagne, etc. (Le Monde)

– on donne une explication, une information supplémentaire :

Le bilan de ces affrontements est très lourd : officiellement 159 morts, surtout des jeunes, et des milliers d'arrestations, officieusement, plus de 500 victimes. (Le Monde)

4. Les résultats

Afin de vérifier que le nouveau graphe *Sentence* apporte une amélioration significative par rapport au graphe fourni avec INTEX, nous avons comparé les résultats obtenus sur 4.5 Mo de textes composés d'articles du journal *Le Monde*, du journal *Ouest-France* et du roman de Balzac *La Femme de trente ans*.

Comme les erreurs de découpage de phrases sont principalement dues aux ambiguïtés avec le point, il nous a semblé intéressant de calculer les fréquences des différentes sortes de points afin de comparer plus justement les résultats. Il ressort de cette étude (tableau ci-dessous) que le roman de Balzac contient très peu de motifs qui peuvent être ambigus avec le point final (2 % de points appartiennent à des anthroponymes) ce qui va expliquer la faible différence de résultats entre les deux automates. Dans le journal *Le Monde*, il y a 8 % de points qui appartiennent à des motifs anthroponymiques, 4,5 % à des abréviations et 0 % de sigles (les articles du *Monde* étudiés ici ont une dizaine d'années comme ceux de *Ouest France*, mais ce journal à tirage national avait alors déjà adopté la norme en vigueur aujourd'hui pour l'écriture des sigles). Le journal *Ouest France* contient environ 20 % de points qui ne sont pas des points de fin de phrases et qui peuvent donc être ambigus avec ceux-ci.

	Ouest-France	*Le Monde*	*La Femme de trente ans*
Points de fin de phrases	69 %	79 %	86 %
Points de suspension	9 %	8 %	12 %
Points dans les sigles	13 %	0 %	0 %
Points dans les anthroponymes	5 %	8 %	2 %
Points dans les abréviations	1 %	4,5 %	0 %
Autres points	3 %	0.5 %	0 %

On compte dans les deux textes les silences et les bruits qu'ont provoqués les graphes. On calcule ensuite deux mesures très couramment utilisées en informatique : le rappel et la précision. Le rappel est le nombre de phrases correctes trouvées par le graphe sur le nombre de phrases qui auraient dû être trouvées de manière idéale ; il calcule la proportion de phrases correctement trouvées. La précision représente le nombre de phrases correctement trouvées sur le nombre de phrases correctes et incorrectes trouvées.

Précision	*Ouest France*	*Le Monde*	*La Femme de trente ans*	*Totaux*
graphe sentence actuel	96,66	98,15	99,63	98,07
nouveau graphe sentence	99,58	99,91	99,92	99,76
Rappel				
graphe sentence actuel	96,93	99,04	99,47	98,44
nouveau graphe sentence	99,69	99,96	99,95	99,83

Comme nous l'attendions, le rappel et la précision sont largement améliorés, surtout dans le journal *Ouest France*. Ceci s'explique par la très grande quantité de points non finaux présents dans ce journal et désormais pris en compte. Les résultats du découpage du journal *Le Monde* sont eux aussi améliorés de manière significative. Le roman de Balzac obtient également une amélioration, même si celle-ci est moins nette.

Avec le nouveau graphe, les silences sont quasiment inexistants (silence = 1 – rappel), mais il reste encore un peu de bruits (bruit = 1 – précision).

Méthodologie de vérification du découpage

Au cours de cette étude nous nous sommes rendu compte que les erreurs communes aux deux graphes étaient très rares. Ces erreurs découvertes dans les corpus étudiés sont des bruits de deux types (remarque : nous n'avons trouvé aucun silence commun !) :

1. Un sigle suivi d'un mot commençant une majuscule :

La C.F.D.T.{S} Basse-Normandie lance un mouvement de grève.

2. Des points de suspension suivi d'un mot commençant par une majuscule :

La littérature vendéenne est souvent signée des plus grands noms : Hugo, Balzac, Chateaubriand, Nerval et même ...{S} Jules Verne.

Étant donné le peu d'erreurs communes aux deux graphes, nous proposons de ne regarder que les erreurs non communes pour vérifier les améliorations apportées par le nouvel automate. Ceci permet d'automatiser en partie la vérification des résultats. Il suffit de comparer les deux textes obtenus après découpages des phrases avec les deux graphes. Lorsqu'on trouve, dans un texte, le symbole {S} de fin de phrase et qu'il n'est pas à cette position dans le second texte, on affiche une concordance des deux textes à cet endroit et on propose un menu dans lequel l'utilisateur choisit quelle solution est la bonne : la première, la deuxième, aucune des deux ou les deux. Lorsque le texte est entièrement vérifié, le programme calcule lui-même les statistiques correspondantes (rappel, précision).

– Exemple 1 : la première solution est la bonne.

...canalisation.{S}Paul BODIN. {S} (1) ONIVINS et INAO : Offices des vins...
...canalisation.{S}Paul BODIN. (1) ONIVINS et INAO : Offices des vins...

– Exemple 2 : la deuxième solution est la bonne.

...et mise à prix 400 000 F. Un Marquet de 1921, Le port d'Alger et la ville,...
...et mise à prix 400 000 F.{S} Un Marquet de 1921, Le port d'Alger et la ville,...

– Exemple 3 : les deux solutions sont bonnes mais l'automate de découpage des phrases ne place pas le symbole {S} au même endroit. L'utilisateur ne doit donc pas compter ceci comme une erreur.

...Luxembourgeois, Hollandais.{S}.. On attendait encore des Américains, ...
...Luxembourgeois, Hollandais...{S} On attendait encore des Américains, ...

Ou les deux solutions sont bonnes mais l'automate de découpage des phrases ne place pas le symbole {S} dans un des deux cas. Le placement ou non du symbole {S} dans le cas ci dessous ne relève pas d'une erreur mais du choix fait pour réaliser le découpage des phrases. L'utilisateur ne doit donc pas compter ceci comme une erreur.

...Sa conclusion est double : « Oui, la Révolution a réussi sa déchristianisation...
...Sa conclusion est double :{S} « Oui, la Révolution a réussi sa déchristianisation...

– Exemple 4 : les deux solutions sont fausses mais le {S} n'est pas placé exactement au même endroit (un espace en plus dans la deuxième solution), l'automate de découpage des phrases place un symbole {S} inutile. Ceci doit être compté comme une erreur.

...-Normandie en lien avec F.R.{S} 3 Normandie s'associera au Prix du polar et ...
...-Normandie en lien avec F.R. {S}3 Normandie s'associera au Prix du polar et...

Ce système semi-automatisé permet de faire une vérification plus rapide d'un corpus beaucoup plus grand. Les erreurs communes étant très peu fréquentes il n'est pas très préjudiciable de ne pas les prendre en compte (0 % de silences communs dans les 3 textes étudiés, 0.30 % de bruits communs pour le journal *Ouest France* mais seulement 0.04 % pour *Le Monde* et *La Femme de trente ans*).

5. Conclusion

Nous venons de le voir, il est possible d'améliorer sensiblement les résultats du découpage d'un texte en phrases, en affinant la description du contexte des signes de ponctuation potentiellement marqueurs de fin de phrase. Mais, inévitablement, certains problèmes subsistent tels que :

– l'apparition d'abréviations non recensées dans nos graphes ;
– l'impossibilité de tenir compte actuellement des guillemets ;
– l'ambiguïté posée par les points de suspension suivis d'une majuscule, où l'on ne peut trancher avec une méthode comme la nôtre.

Par ailleurs, la manière dont les graphes sont passés sur le texte pose un problème : si plusieurs cas sont imbriqués, il peut se produire une erreur. L'exemple suivant illustre ce problème.

Ce produit est vendu 20 F. {S} Tél.{S} : 03.25.00.01.02.03

Le motif *20 F. Tél* est reconnu dans le graphe cas3 par le chemin motifSymbole suivi d'un point, d'un espace et d'un mot commençant par une majuscule. Cette séquence ayant été reconnue, la reconnaissance des motifs redémarre après le mot *Tél.* Le motif *Tél. : 03.25.00.01.02.*03 (décrit dans le graphe *TelFax*) n'est alors pas reconnu et la fin de phrase est mal placée. Cette imbrication ne peut être prise en compte car elle complique énormément le graphe et risque d'induire d'autres erreurs (plusieurs chemins peuvent avoir la même longueur auquel cas l'algorithme d'INTEX ne choisit pas forcément le chemin que l'on souhaite).

De manière générale, dans les résultats de notre découpage, nous avons privilégié le silence au bruit, sauf dans la mesure où le silence aurait été beaucoup plus fréquent que le bruit (c'est le cas des points de suspension).

S'il n'est pas possible de rendre infaillible le découpage automatique des phrases, on peut encore l'améliorer sur un certain nombre de points, notamment par le recensement systématique, dans des domaines précis, des mots composés utilisant des lettres ou encore des abréviations d'usage.

Bibliographie

DISTER (Anne) : 1997, « Problématique des fins de phrase en traitement automatique du français », in *À qui appartient la ponctuation? Actes du colloque international et interdisciplinaire de Liège (13-15 mars 1997)* (Bruxelles : Duculot, « Champs Linguistiques »).

GREFENSTETTE (Gregory), TAPANAINEN (Pasi) : 1994, « What is a word, what is a sentence? Problems of Tokenization », in *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)* (Budapest).

RICHAUDEAU (François) : 1989, Manuel de typographie et de mise en page (Paris : Éditions Retz).

DRILLON (Jacques) :1991, *Traité de ponctuation française* (Paris : Gallimard).

SILBERZTEIN (Max) : 1993, *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX* (Paris : Masson).

SILBERZTEIN (Max) : 1998, « Normalisation des textes », in *Actes des Troisièmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1998)* (Nice), p. 601–614.