DEFINITION DE CRITERES FORMELS DANS L'USAGE DU DEUX-POINTS POUR LE TRAITEMENT AUTOMATIQUE DE LA LANGUE NATURELLE

M. AKBAL

Abstract

How to make a computer understand a written text ? This is the problematic ofthe automatic processing of the language, a division ofthe artificial intelligence. It is based on simulation and recogning. The process consists offinding formai criteria by dealing superficially with the language. In the text below, we shall try to identify those which concern the usage ofthe colon.

Introduction

L'objectif de la présente étude, dont l'originalité demeure entière, est de contribuer à trouver un angle d'attaque en mesure de définir les bases formelles dans l'utilisation du deux-points.

Ces bases formelles sont nécessaires dans le cadre du traitement automatique de la langue naturelle pour deux raisons essentielles :

1 - La ponctuation joue un rôle important dans la logique du discours et sa compréhension. "C'est la ponctuation qui apporte la lumière, permet la certitude et la précision du message" (Doppagne, A. 1984).

2 - Le deux-points est un signe assez particulier. "Le deux-points comme la reine des échecs, peut marcher en avant, en arrière et en diagonale" écrit Drillon. "Le deux-points symbolise à la perfection l'ambiguïté de l'ergo cartésien" poursuit-il (Drillon, J. ; 1991).

1-Ponctuation et traitement automatique de la langue

L'homme interprète, déchiffre, donne sens et dit l'esprit caché du discours et de la langue. C'est le privilège de l'homme, dira-t-on. Mais le pouvoir de donner sens au langage écrit ou parlé trouve sa véritable limite dans le monde de l'ordinateur. L'automate est incapable de concevoir des grilles d'interprétation lui permettant d'expliquer les textes. Véritable leçon d'impuissance... Ceci trouve sa raison dans le fait que l'objet à traiter, déformer, refomer et restituer est un système dynamique, flexible, interactif et d'une déroutante relativité. A cela, il convient d'ajouter la structure limitative de l'automate qui, à l'état actuel de l'art, ne possède pas cette capacité d'accomplir des raisonnements par induction : basé sur la simulation et le calcul.

1-1- Les analyseurs morpho-syntaxiques du français écrit

La méthode du traitement automatique de la langue naturelle consiste à définir avant le modèle algorithmique le modèle linguistique. Et la conception du modèle linguistique, fondé sur le traitement en surface de la langue, passe par quatre niveaux d'analyse : morphologique, syntaxique, sémantique et pragmatique.

1 - Morphologique : Il s'agit de renconnaître les formes et d'en déduire les entrées lexicales ou lemmes. Cette étape est aujourd'hui plus ou moins résolue dans le cas de l'écrit. Mais là où le bât blesse c'est lorsque l'on sait que la langue française compte environ un million de noms propres non encore normalisés. Auxquels viennent s'ajouter les abréviations usuelles et les mots étrangers.

2 - Syntaxique : A partir des formes, reconstruire la phrase. A chaque forme reconnue est associée sa catégorie grammaticale, en vue d'accéder à la structure syntaxique de la phrase. Cette étape, qui quoique n'est pas facilitée par les formes auxquelles on peut associer plus d'une catégorie grammaticale, est réglée.

3 - Sémantique : Passer de la phrase à sa signification. Cette étape n'est pas résolue. Bon nombre de problèmes se posent à ce niveau :
- Les ambiguïtés morphologiques, exemple : "La petite brise la glace" a deux interprétations ; seul le contexte permet de choisir.
- L'imprécision des attachements prépositionnels (les ambiguïtés syntaxiques), exemple : "Je mange du couscous avec des amis". Dans cette phrase, on ne sait pas si le syntagme prépositionnel est rattaché à "je" ou à "couscous".
- Les ambiguïtés sémantiques, exemple : le verbe "voler".
- La difficulté de détecter les néologismes : le verbe "zapper"
- La difficulté de détecter les expressions imagées : "à couteaux tirés".

4 - Pragmatique : Elle consiste à préciser les limites et les motivations de la phrase selon le fond culturel, la connaissance du monde. C'est un problème vaste qui en est à ses premiers balbutiements. En effet, le scripteur et le lecteur partagent une connaissance du monde non véhiculée par les messages. A la question "Peux-tu me dire l'heure?", au lieu de répondre" Il est 23h00", un ordinateur dira "Oui".

1-2- La ponctuation et les analyseurs morpho-syntaxiques

Pans les analyseurs morpho-syntaxiques de l'écrit, qui ont été jusque-là conçus, les ponctuations sont regroupées dans une catégorie, et sont sous-catégorisées selon la valeur d'une variable qui marque la force de la ponctuation.
VI : Ponctuations fortes (marqueurs de fin de phrases)

. ! ?

V2 : Ponctuations faibles (marqueurs de frontières de syntagmes)

V3 : Ponctuations balancées

( ) " " { } [ ] « »

Les ponctuations ne sont utilisées par ces systèmes que pour leur valeur démarcative. Deux critiques peuvent être formulées :
1- Nous notons d'abord que les valeurs démarcatives affectées à chacune des trois sous-catégories ne sont pas toujours vraies. Citons le point d'interrogation et le point d'exclamation qui peuvent ne pas être des marqueurs de fin de phrases. Ajoutons le cas du point virgule qui n'est pas une frontière de syntagme.

2 - Les ponctuations ont d'autres valeurs autre que démarcatives. Le deux-points apporte des informations supplémentaires qui ne sont pas, jusque là, exploitées par ces systèmes.

A cela, il convient d'ajouter que la ponctuation pose certaines ambiguïtés au niveau reconnaissance des formes qu'on ne sait lever, actuellement, que manuellement. Citons à titre d'exemple : l'ambiguïté posée par le point (point de fin de phrase vs point d'abréviation), celle posée par le tiret (tiret vs trait d'union) et celle de la majuscule (rôle démarcatif ou distinctif).

2 - Problématique

A ce stade de notre propos, disons que les ponctuations posent des problèmes dans le traitement automatique de l'écrit.
Dans cette optique, le 'deux-points se distingue des autres signes de la ponctuation au point qu'il laisse perplexe J. Gracq. "Dans le groupe de signes de la ponctuation, il en est un qui n'est pas tout à fait de la même nature que les autres : les deux-points. Ni tout à fait ponctuation, ni tout à fait conjonction, il y a longtemps qu'il me pose des problèmes d'écriture. Tous les autres signes, plus ou moins marquent des césures dans le rythme, ou des flexions dans le ton de la voix ; il n'en est aucun, sauf lui, que la lecture à voix haute ne puisse rendre acceptablement. Mais dans les deux-points, s'embusque une fonction autre, une fonction active d'élimination; ils marquent la place d'un mini-effondrement dans le discours, effondrement où une formule conjonctive surnuméraire a disparu corps et biens pour assurer aux deux membres de phrases qu'elle reliait un contact plus dynamique et comme électrisé: il y a toujours dans l'emploi des deux-points la trace d'un menu court-circuit. Ils marquent aussi, à l'intérieur du discours lié, un début de transgression du style télégraphique; une étude statistique révélerait sans doute le peu d'usage qu'en ont fait les auteurs anciens (jusqu'où d'ailleurs son usage remonte-t-il ?) tout comme sa fréquence grandissante dans les textes modernes. Tout style impatient, soucieux de rapidité, tout style qui tend à faire sauter les chaînons intermédiaires, a spécialement affaire à lui comme un économiseur, péremptoire et expéditif ' (cité par Bessonet, D. ; 1991).

A la lumière de cette citation, qui illustre amplement notre problématique, quatre idées maîtresses peuvent

être retenues :
- la ponctuation est étroitement liée à la prosodie,
- le deux-points est un outil de concision,
- le deux-points est un signe révélateur de style,
- le deux-points renferme la nature de contact entre les membres reliés de la phrase.

Drillon précise que "sa signification actuelle est fort récente puisqu'elle ne remonte qu'au XIXe siècle. Auparavant, il avait valeur de ponctuation forte, supérieure au point virgule" (Drillon, J. ; 1991).

Pour notre part, nous émettons l'hypothèse que le deux-points laisserait l'automate impuissant pour trois raisons :
- La diversité des fonctions et des places qu'il occupe dans les combinaisons.
- Le caractère diversifié des rôles syntactico et logico-sémantique qu'il peut prendre.
- Et enfin, sa valeur pragmatique évasive, c'est-à-dire difficile à cerner et à maîtriser.

3 - Fonctions et places du deux-points

3-1- Le deux-points : un connecteur

D'après Drillon, le deux-points joue un "rôle de simulateur logique et chronologique" (Drillon, J. ; 1991). Simulateur logique : il fait office de connecteur. Simulateur chronologique : il marque le passage entre ce qui précède et ce qui suit.

Le deux-points relie : deux ou plusieurs propositions, une proposition principale et un ou plusieurs syntagmes nominaux, ou plusieurs propositions et plusieurs syntagmes nominaux.

C'est un marqueur argumentatif qui ne se précise que dans le contexte de son apparition, il explique, commente, justifie, expose, montre, prouve, rend compte, traduit, dit, ...

Il a aussi le pouvoir d'énumérer, d'inventorier, de classer, de distinguer, de répertorier, de dénombrer,

Comme les autres signes de la ponctuation, il sert à lever les ambiguïtés dans le discours. C'est un séparateur de propositions et/ou de syntagmes. Il provoque dans le texte une digression, une cassure doublement incitatrice.

Primo, inciter le lecteur à s'attarder sur un passage donné, sur une idée, sur un paradigme parce qu'ils présentent un quelconque intérêt pour la compréhension de l'énoncé lui même ou du texte. A titre illustratif, dans la phrase suivante : "Cependant, un second facteur intervient pour sa compréhension : le contexte de son apparition", le,deux-points sert à mettre en relief le syntagme "le contexte de son apparition". La substitution du deux-points par le pronom démonstratif "ce" et du verbe "être" conjugué à la troisième personne du présent de l'indicatif n'affecterait pas le sens de la pharase. Toutefois, l'usage de ceci ou de cela ne produirait pas le même effet de lecture. Dans le premier usage, le lecteur est sujet à un arrêt provoqué par le deux-points. Dans le deuxième, par contre, la phrase risque de filer sans qu'il n'ait à le marquer. C'est le "menu court-circuit" de Gracq.

Secundo, le deux-points est utilisé pour une meilleure clarté du texte. Il permet au lecteur de saisir et de comprendre sans effort le discours. L'exemple le plus illustratif se situe au niveau du discours énumératif dans lequel le scripteur fait l'inventaire d'une série d'éléments qu'il organise selon des critères logiques.

En introduisant une citation, le deux-points discrimine judicieusement ce qu'on emprunte aux autres et le produit de notre pensée.

Aussi, le deux-points remplace un discours elliptique que l'auteur se refuse de faire figurer dans la phrase: soit "pour se soustraire (...) aux dures nécessités de la phrase complexe, ou tout simplement parce que ce qui est supprimé figure déjà dans le texte" (Drillon, J. ; 1991).

Il nous est possible de considérer que la redondance, dans l'emploi du deux-points et du connecteur qu'il remplace, provoquerait de l'effet pour le lecteur dans la mesure où le deux-points l'inciterait à marquer une pause et le connecteur lui permettrait de connaître le type de relation qu'il y a entre les énoncés de gauche et ceux de droite.

3-2- Le deux-points : rôle démarcatif

"La phrase commence par une majuscule et se termine par un point". Cette règle grammaticale classique dénote que chaque phrase comporte un marqueur de début de phrase (MDP) : une majuscule (nous supposons résolu le problème des noms dans lesquels la majuscule a un rôle distinctif), et un marqueur de fin de phrase (MFP) : un signe de la

ponctuation relevant de la sous-catégorie des ponctuations fortes (VI . ? !).

Il importe de signaler que ce schéma n'est pas toujours respecté. Certaines phrases ont une structure formelle tout autre.

Trois cas de figure qui n'obéissent pas au schéma classique peuvent être retenus.
Cas 1.
MDP : "MDP "MFP
Dans ce cas, le deux-points relève de la catégorie VI (MFP). Elle est facile à reconnaître formellement par le biais des guillemets.
Cas 2.
MDP:
- MDP MFP
-MDP MFP
- MDP MFP
Dans ce schéma le deux-points est une ponctuation forte, sa portée dépasse largement le cadre de la phrase.
Cas3.
MDP:
-m ;
; - m MFP
M MFP
-m MFP

Ce schéma est insolite. Une phrase est emboîtée dans une autre. Ces cas de figures remettent en "cause l'analyse actuelle des analyseurs morpho-syntaxiquess qui considère que le deux-points relève de la sous-catégorie V2 (ponctuations faibles). Il nous est permis d'affirmer que le deux-points est la seule ponctuation dont le rôle démarcatif est aussi fluctuant.

3-3- Le deux-points : un signe polysémique

Le deux-points est le seul signe de la ponctuation qui est polysémique. L'idée de polysémie devient intéressante dans la mesure où le deux-points n'est pas limité à un seul sens et donc, à un seul type de structure syntaxique. Selon son occurrence le deux-points s'ajuste au contexte.

Il peut commuter avec : certaines conjonctions de coordination, certaines conjonctions de subordination et certaines prépositions.

3-3-1- Le deux-points et les conjonctions de coordination

Dans certains contextes le deux-points peut être remplacé par certaines conjonctions de coordination : "mais", "ou", "donc", "et", "or", "ni", "car".

Une distinction formelle est à faire entre ces conjonctions. En effet, seules "et", "ou" et "ni" peuvent coordonner à la fois des syntagmes et des propositions ; les autres coordonnent en général des propositions entre elles. Toutefois, ce critère ne peut à lui seul, constituer une base formelle régulatrice.

Aussi, ces conjonctions sont des connecteurs logiques : "mais" exprime une opposition, "ou" un choix, "or" l'exhortation, "donc" une conclusion, "et" une addition, une intersection, une liaison voire un rapprochement et "car" introduit une explication.

Dans ces cas de figures, il n'existe pas de cadres formels significatifs, dans l'usage du deux-points, qui seraient d'un quelconque apport au traitement de la langue.

3-3-2- Le deux-points et les conjonctions de subordination

Dans d'autres contextes, le deux-points commute avec : la conjonction de subordination "parce que". Il peut aussi être remplacé par les prépositions suivantes : "pour", "afin de", "en vue de", "dans le but de", "dans l'intention de",.-- Ces prépositions expriment le but. Un seul critère formel, sous réserve de vérifier sa régularité, serait l'usage de l'infinitif juste après le deux-points.

Par surcroît, il existe d'autres conjonctions qui introduisent des subordonnées circonstancielles de but : "pour que", "afin que", "de sorte que", "de peur que",... Elles se distinguent des précédentes au niveau de la structure de la phrase. Probablement, un critère formel : le verbe de la subordonnée se met au subjonctif. L'emploi du subjonctif nous laisse aussi supposer que le deux-points est utilisé comme substitut de la conjonction de subordination "bien que". Pouvons nous retenir son usage comme critère formel qui rend suffisamment compte de cette commutation ? Est-il possible d'étendre ce critère à la conjonction "quoi que" qui exprime une opposition ou une concession ?

3-4- Les relations logico-sémantiques

En plus des relations logiques citées supra, le deux-points exprime d'autres relations : d'allotaxie, métalinguistique et d'hyperonymie/hyponymie vs hyponymie/hyperonymie.

3-4-1- Relation d'allotaxie

"L'allotaxie signifie le fait que des formes de phrases différentes permettent d'exprimer la même idée" (Carré, R. et al ; 1991). En d'autres termes, ce qui vient avant le deux-points signifie exactement ce qui vient après, mais exprimé sous des formes différentes.

$$P(x) = P(x')$$

Il est difficile de dégager des critères formels généraux susceptibles de rendre compte de cette relation, exception faite du discours traduit : présence donc d'une forme inconnue.

3-4-2- Relation métalinguistique

C'est une relation dans laquelle "la plupart des énoncés comportent, implicitement ou explicitement, une référence à leur propre code" (Ducrot, 0. & Todorov, T. ; 1972).

La présence d'un critère lexical à gauche du deux-points ("définition") est-il suffisamment pertinent pour rendre compte d'une telle relation ?

3-4-3- Relation d'hyperonymie/hyponymie vs hyponymie/hyperonymie

"Un hyperonyme (est un) terme dont le sens inclut le sens d'autres termes qui sont ses hyponymes" (Carré, R. étal. ; 1991).
Cette relation se présente sous la forme suivante:

$$P(x) = x1, x2, x3, x".$$

C'est-à-dire que nous trouvons à droite du deux-points un hyperonyme et à sa gauche une énumération d'hyponymes.

En revanche, dans la relation dite d'hyponymie/hyperonymie, les hyponymes se trouvent avant le deux-points et l'hyperonyme après le deux-points. Cette relation peut être formulée comme suit :

$$x1, x2, x3, x" \bullet P(x).$$

Est-ce-que les virgules qui séparent les syntagmes forment un critère suffisant ?

3-5- Bilan partiel

Dans ce qui précède, nous avaons montré que le deux-points joue un rôle de séparateur et/ou de connecteur dans les combinaisons.

C'est aussi un signe syntaxiquement ambigu dont les sens ne se précise que dans le contexte de son apparition. Il est soit précise que dans le contexte de son apparition. Il est soit conjonction de coordination, soit conjonction de subordination ou préposition. Il véhicule de nombreuses relations logico-sémantiques.

Cet état de fait ne nous a pas permis de trouver des bases formelles nécessaires à la prise en compte du deux-points dans le cadre du traitement automatique du français écrit.

4- Un autre angle d'attaque : les types de discours introduits par le deux-points

4-1- Le discours explicatif

Le discours explicatif implique l'existence au moins de deux propositions juxtaposées (une proposition principale et au moins une proposition explicative) jointes par un deux-points jouant le rôle de lien de coordination ou qui lui est proche.

Le deux-points qui introduit le discours explicatif peut être remplacé par une panoplie de locutions et d'expressions. Dans cet optique, il faut citer la conjonction de coordination : "car". A cela, il convient d'ajouter d'autres expressions explicites : "s'explique par", "c'est-à-dire que", ...

Le deux-points a aussi ce pouvoir logique, en ce sens où il peut commuter avec certaines locutions telles que : "parce que", "à cause de", "a pour conséquence",...

Dans certains cas du discours explicatif, remarquons le phénomène de l'anaphore. En effet, ce qui vient après le deux-points commence dès fois soit par la reprise, totale ou partielle, d'un syntagme nominal, soit par un pronom anaphorique. Pouvons-nous considérer ce critère comme base suffisante ?

Dans d'autres cas, relevons la présence de verbes ou de locutions qui introduisent une explication avant

ou après le deux-points : "de la façon suivante", "de la manière suivante", ou "il s'agit de", "il correspond à",...

En plus, le deux-points sert :
- à montrer comme avec un geste d'indication les idées, les objets et les êtres impliqués dans le discours, le raisonnement ou l'argumentation,
- à noter simplement qu'ils sont connus du lecteur pour diverses raisons ou parce qu'ils sont cités préalablement dans le texte.

Dans ce contexte, ceux, celles, celui, ce sont, c'est, celui-ci, celle-là, celle-ci, ceux-ci,... sont les expressions qui peuvent remplacer le deux-points ou qui sont placés juste après son apparition. Une caractéristique : tout ceci fonctionne comme des anaphores. Ces démonstratifs sont des anaphoriques et non des déictiques.

4-2- Le discours énumératif

Le discours énumératif traduit l'action d'énumérer ou d'inventorier un ensemble de sous éléments appartenant à un élément principal ou clé cité et contenu dans la ou les propositions qui précédent le deux-points.

Dans ce cas de figure, ce type de discours est introduit par une multitude d'expressions et de verbes : "sont le (s) suivant (es)", "qui suivent", "à savoir", "citer", "énumérer", "classer", "inventorier", "catégoriser",...

Il existe plusieurs critères formels qui rendent compte du discours énumératif. Certains sont utilisés d'une façon singulière et d'autres d'une façon concomitante.

A gauche du /:/ :
- Les déterminants numéraux cardinaux, exemple "deux échelles", "six épreuves", "cinq éléments",...
- un déterminant indéfini servant à la quantification, "plusieurs", "quelques",...
- l'abréviation MM. fait attendre une énumération,
- un critère relevant de la syntaxe : un mot cité à gauche serra répété à droite autant de fois que durera l'énumération,

A droite du /:/ :
- la majuscule pour chaque élément énuméré,
- des syntagmes nominaux séparés par des virgules,
- une conjonction de coordination "et" qui indroduit le dernier syntagme,
- les caractères numériques, alphabétiques ou mathématiques (système de numérotation),
- le passage à la ligne entre chaque élément de l'énumération,
- le tiret,
- l'italique,
- présence de syntagmes nominaux dont le centre est un déverbal,
- des structures lexicales : "d'un côté, de l'autre", "d'une part, et d'autre part", "de l'extérieur, de l'intérieur", "premièrement, deuxièmement", "primo, secundo", "l'un et l'autre",
- la conjonction disjonctive "ou", elle indique une alternative,
- la redondance de la conjonction de coordination "ni".

4-3- Le discours illustratif

Le discours exemplificatif quand il vient après le deux-points sert à illustrer une ou les idées contenues dans la ou les propositions qui précédent le deux-points par un ou des exemples concrets.

Les expressions suivantes peuvent être retenues pour le remplacement éventuel du deux-points : "comme", "tel (les) que", "à titre d'exemple", "par exemple", ...

Il faut noter, qu'hormis certains indices lexicaux, il est difficile de trouver des critères formels qui rendent compte de de type de discours.

4-4- Bilan partiel

Nous avons relevé que le deux-points introduit généralement trois types de discours : explicatif, énumératif et illustratif.

Pour le discours énumératif, nous sommes arrivés à dégager des critères formels pouvant rendre éventuellement compte d'un modèle d'organisation discursive assez homogène et plus ou moins rigide.

Par contre, les discours explicatif et illustratif sont d'une organisation lâche laissée au gré des fluctuations énonciativos des scripteurs.

5- Conclusion générale

A la fin de ce travail, nous ne prétendons pas avoir épuisé le sujet, il s'agit d'une réflexion préliminaire pour d'ultérieures investigations plus fines. Car le sujet de par sa complexité et la richesse des aspects qui le recouvrent demeure toujours passionnant et garde tout son intérêt et sa nouveauté.

Ainsi, d'ores et déjà, en ayant pris conscience de l'étendue du problème, il nous est possible d'émettre un certain nombre d'interrogations pouvant servir de pistes de recherche :

- Il peut être admis que le deux-points introduit bon nombre de relations logico-sémantiques: métalinguistiques, d'allotaxie, d'hyperonymie/hyponymie VS d'hypony mie/ hyperonymie, logiques, définitionnelles, d'ingrédience,... Il reste à savoir dans quelle mesure ces relations peuvent faire l'objet d'une formalisation ?

- Peut-on affirmer qu'un énoncé qui suit le deux-points est (toujours? souvent? dans des cas qu'on peut identifier de préférence formellement ?) un énoncé à haute valeur informative parce qu'il annonce de façon synthétique une idée importante du document? Dans ce cas, le deux-points pourrait jouer un rôle de signal qu'on pourait paraphraser en : "je vais dire quelque chose d'important".

- Partant de l'idée que le rôle démarcatif du deux-points est fluctuant, en ce sens qu'il peut être tantôt dans la sous-catégorie des ponctuations fortes tantôt dans la sous-catégorie des ponctuations faibles, n'est-il pas tentant de revenir sur la classification existante pour d'éventuels réaménagements?

6- Bibliographie

- Bessonet, D. ; Enseigner la ... "ponctuation"?!. In Pratiques, n°70, juin 1991.

- Carré, R. & Dégrémont, J. F. & Gross, M. & Pierrel, J. M. & Sabah, G. ; Langage humain et machine, Paris, Presses du CNRS, 1991.

- Doppagne, A. ; La bonne ponctuation : clarté, précision, efficacité de vos phrases, 2ème éd., Paris, Ed. Duculot, 1984.

- Drillon, J. ; Traité de la ponctuation française, Paris, Gallimard, 1991.

- Ducrot, O. & Todorov, T. ; Dictionnaire encyclopédique des sciences du langage, Paris, Seuil, 1972.