Les structures syntaxiques du langage Vercingétorix I. Problèmes liés à leur naturalité.

Gabriel G. Bès

Les structures syntaxiques du langage Vercingétorix I. Problèmes liés à leur naturalité.

Gabriel G. Bès
Groupe de recherches sur la condensation de l'information en langue naturelle (CILN)
Université Blaise-Pascal, Clermont II

Traitement automatique des langues naturelles et systèmes documentaires, Condenser, Supplément n° 1, Adosa, Clermont-Ferrand, 1984, p. 109-134.

Résumé

Le langage documentaire Vercingétorix 1 (V1) est un outil pour la recherche documentaire, destiné à être utilisé par le documentaliste indexeur pour produire des indexations et les questions documentaires. Pour satisfaire diverses exigences, on a fait, dans la formulation de V1, le choix de la naturalité. Ce texte attire l'attention sur les désavantages de cette naturalité avec deux objectifs de fond : a) apporter des éléments pour un bilan critique de V1 sur le plan technologique et b) dégager les caractéristiques des langues naturelles susceptibles d'expliquer ce bilan. La discussion s'articule autour des notions de paraphrase et ambiguïté, caractéristiques des langues naturelles et sur les compromis à faire dans le langage V1 (ajout ou non de règles d'utilisation susceptibles de complexifier excessivement le système).

Voir aussi

Gabriel G. Bès et Pierre-Maurice Fauchère. « Le système documentaire Vercingétorix I ». *Condenser*, Adosa, Clermont-Ferrand, février 1980, n° 1, p. 57-94. http://hal.archives-ouvertes.fr/hal-01071423

Gabriel G. Bès et Pierre-Maurice Fauchère. « Le langage V1 : structures engendrées par la grammaire et relations entre structures grammaticales (1re partie) ». *Condenser*, Adosa, Clermont-Ferrand, janvier 1981, n° 2, p. 39-89. http://hal.archives-ouvertes.fr/hal-01117883

Gabriel G. Bès et Pierre-Maurice Fauchère. « Le langage V1 : structures engendrées par la grammaire et relations entre structures grammaticales (2e partie) ». *Condenser*, Adosa, Clermont-Ferrand, avril 1982, n° 3, p. 3-31. http://hal.archives-ouvertes.fr/hal-01117888

NB. La version 2 de ce document déposée dans HAL ne diffère de la version 1 que par les liens ci-dessus, qui ont dus être corrigés suite à une défaillance du système ayant conduit au remplacement de certains dépôts par d'autres n'ayant rien à voir.

Gabriel G. Bès

LES STRUCTURES SYNTAXIQUES DU LANGAGE VERCINGÉTORIX I. PROBLÈMES LIÉS A LEUR NATURALITÉ

Le langage V-1. Ses objectifs et conditions d'adéquation

Le langage documentaire Vercingétorix 1 (désormais V-1) est un outil pour la recherche documentaire. V-1 porte aujourd'hui sur le domaine des sciences de l'information, mais son schéma général d'organisation devrait pouvoir s'adapter à d'autres domaines. Comme pour Syntol ou Précis, l'indexeur doit s'en servir pour produire des indexations et les questions documentaires. La machine doit, par la suite, traiter ces questions dans la recherche documentaire et fournir les réponses. En tant que langage documentaire, sa condition d'adéquation la plus générale sera sa capacité de réduire le silence et le bruit dans la recherche documentaire. A cette condition générale on a ajouté les suivantes :

(a) Acquisition rapide et utilisation efficace par le documentaliste des règles du langage qui conditionnent son travail d'indexation. On exige en particulier l'uniformité des indexations, c'est-à-dire que des documentalistes différents opérant sur un même document, interprété de la même manière, aboutissent à un même résultat.

(b) Double utilisation des indexations : 1) traitement automatique pour la recherche documentaire ; 2) lisibilité humaine afin de produire des outils bibliographiques tels que bulletins bibliographiques de résumés.

(c) Adaptabilité du schéma général du langage aux exigences de chaque situation particulière. En particulier, les discriminations introduites par les relations syntaxiques s'inscrivent dans une échelle de complexité croissante entre un niveau 0, qui correspond à un thesaurus sans syntaxe, et un niveau n, qui correspond à la syntaxe d'une langue naturelle (désormais LN). Chacun des niveaux intermédiaires doit incorporer les discriminations au niveau précédent et en ajouter d'autres. Le "1" de V-1 note un de ces niveaux. Cette caractéristique doit permettre aussi d'affiner les indexations d'un centre de documentation sans perte du fond indexé selon le niveau précédent.

* Je remercie M. CHAMBREUIL et P.-M. FAUCHERE des utiles discussions dans la préparation de ce travail.

Traitement automatique des langues naturelles et systèmes documentaires, « *Condenser.* Supplément 1 », Clermont-Ferrand, 1984.

(d) Utilisation du langage dans un système multilingue, avec non seulement possibilité d'indexer des documents écrits en langues différentes, mais aussi de produire des indexations et des interrogations qui seront différentes selon la langue du documentaliste.

(e) Coût informatique "modéré", aussi bien dans la phase de mise au point d'un langage particulier, que dans la phase de ppoduction des indexations et de la recherche documentaire.

Pour satisfaire les conditions précédentes, on a fait, dans la formulation de V-1, le choix de la naturalité. Dans une première approximation, on dira que, en imposant un certain nombre de caractéristiques au système sous-jacent de V-1, on veut s'approcher de la situation considérée comme idéale dans laquelle le coût d'acquisition et d'utilisation du langage serait très faible pour le documentaliste : dans cette situation limite, il devrait pouvoir écrire avec son langage documentaire comme s'il écrivait dans sa propre langue.

Le thème de la communication n'est pas la présentation des avantages de la naturalité mais, plutôt, d'attirer l'attention sur les désavantages, avec deux objectifs de fond : a) apporter des éléments pour un bilan critique de V-1 sur le plan technologique ; b) dégager les caractéristiques des LN susceptibles d'expliquer ce bilan, que celui-ci soit positif ou négatif.

Rappel de la composition de V-1

On ne donne ici qu'une présentation sommaire de V-1 qui devrait perpettre de suivre la discussion ci-dessous ; pour une présentation détaillée, cf. G.G. Bès et P.-M. Fauchère, Condenser 1, p. 57-94, Condenser 2, p. 41-89, Condenser 3, p. 5-31.

Soit la suite complète suivante, produite par le documentaliste utilisant V-1 :
(1) Traduction t1 de documents t0 portant sur s la coopération a1 b internationale a1.

On doit reconnaître en relation à une suite telle que celle-ci plusieurs types d'information :

(a) L'information de base : indicateurs de procédé (dans la suite (1) : t1, t0, a1), le marqueur b et les notions (notées en (2) en capitales).

(b) les articulations grammaticales ; dans la suite : portant sur s.

(c) les emplacements des articulations libres (notées par "..." dans (2)).

(d) les articulations libres remplies par le documentaliste (comparer (2) avec (1)).

(e) l'emplacement du symbole x, qui sera remplacé par une notion particulière.

(f) le système de relations associé à l'information de base.

A partir de ces éléments (que le documentaliste ne doit pas distinguer dans l'écriture de (1)), on peut introduire les distinctions suivantes :

- Suite complète : information de base plus articulations grammaticales plus articulations libres remplies par le documentaliste ; ex. : cf. (1).

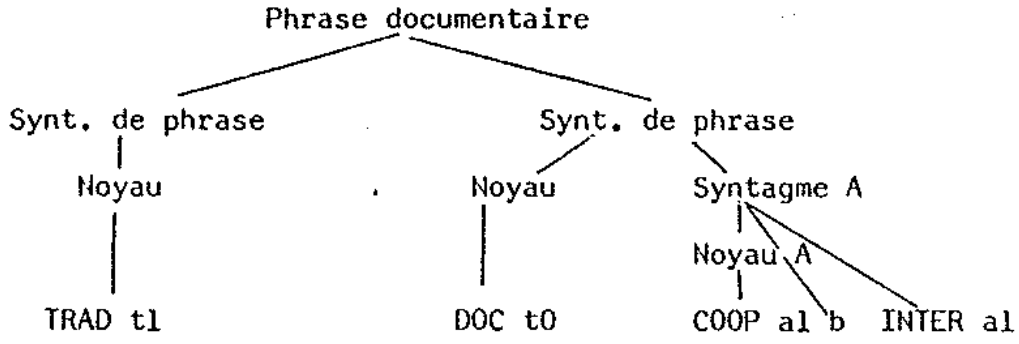
- Suite individuelle (désormais suite indiv.) : information de base plus articulations grammaticales plus les emplacements des articulations libres :

(2) ... TRAD ... t1 ... DOC ... t0 portant sur s ... COOP ... a1 b ... INTER ... a1

- Suite syntagmatique (désormais suite syntag.) : indicateurs de procédé plus le marqueur b plus les emplacements du symbole x plus les articulations grammaticales plus les emplacements des articulations libres ; ex. :

(3) ... x ... t1 ... x ... t0 portant sur s ... x ... a1 b ... x ... a1

Le système de relations associé à l'information de base est le suivant :



Une suite indiv. avec le système de relations qui lui est associé est une structure grammaticale schématique. On parlera parfois de suites partiellement syntagmatiques pour faire référence à des suites où seulement quelques-uns des symboles "x" ont été remplacés par des notions.

La grammaire de V-1 est constituée d'un lexique et d'un ensemble de

règles grammaticales. La grammaire spécifie les structures grammaticales schématiques.

Le lexique de V-1 est constitué d'un ensemble d'entrées lexicales. Chaque entrée lexicale est constituée d'une notion associée à un radical discriminant (noté en capitales dans (2))et d'une série d'indications sur l'utilisation de l'entrée. Pour produire les indexations, le documentaliste doit avoir accès à seulement une partie de ces indications : celles-ci portent notamment sur les indicateurs de procédé admis par chaque notion et sur le mot de rappel en LN associé à chaque notion. Ces indications sont déductibles, pour chaque entrée, de quelques règles de correspondance très générales.

Les indicateurs de procédé permettent un classement syntagmatique des notions. La catégorie notée T comprend les indicateurs individuels t0, t1, t2, t3 ; la catégorie notée TV comprend t1, t2, et t3. La catégorie notée A comprend les indicateurs individuels a1, a2, a3 et a4. Le documentaliste ne note que les indicateurs individuels (T, TV et A ne sont utilisés que pour exprimer les règles du langage).

Les règles grammaticales essentielles de V-1 sont les suivantes :

- Un item est constitué d'une notion notée par un radical discriminant et suivie d'un indicateur de procédé. Les emplacements des articulations libres se situent avant chaque notion et entre chaque notion et l'indicateur de procédé.

- Un syntagme de phrase est constitué par un item T (c'est-à-dire d'une notion suivie d'un des indicateurs de procédé qui appartient à la catégorie T) suivi, optionnellement, par un groupe A (suite d'items A). L'item T est le noyau du syntagme de phrase. Il y a des noyaux simples (T) et des noyaux avec coordination interne (T et T). Le marqueur "b" peut être utilisé entre deux items A : il indique que les items A qui suivent, non précédés d'un autre marqueur b, portent sur l'item A qui précède. Un item A peut être précédé optionnellement d'un spécificateur A (articulation grammaticale) : celui-ci indiquera le type de relation entre l'item A et le noyau sur lequel il porte.

- Un syntagme de phrase dont le noyau est un item TV peut être suivi d'un autre syntagme de phrase. Un ou plusieurs syntagmes de phrase qui se suivent dans une même suite constituent une phrase documentaire.

- Le nombre d'items A dans un syntagme de phrase et de syntagmes de phrase dans une phrase documentaire n'est pas limité.

- Une construction documentaire est constituée de deux phrases documentaires mises en relation par un prédicat suivi du symbole "r".

Les règles textuelles de V-1 permettent de classer les structures grammaticales en deux types : les expressions documentaires (ED) et les sous-expressions documentaires (SED) et introduisent l'anaphorique. Les ED sont les structures grammaticales qui représentent des aspects indépendants du contenu d'un document et justifient l'inclusion d'un document dans une réponse à une question documentaire. Les SED s'intègrent dans une ED.

L'anaphorique -noté essentiellement par les formes ce(s) cette(s) placées devant le terme repris- permet de réintroduire dans une structure grammaticale le contenu significatif d'une structure grammaticale précédente. La discussion de ce travail pouvant être suivie sans recours aux règles qui conditionnent l'utilisation de l'anaphorique, on ne les résumera pas ici.

Les structures spécifiées par la grammaire de V-1 sont en correspondance avec des suites en LN. Ces correspondances se définissent à plusieurs niveaux (cf. ci-dessous). A un de ces niveaux on trouve les correspondances entre suites syntagmatiques V-1 et suites syntagmatiques en LN. (Par suite syntagmatique LN on comprend la disposition linéaire en surface des catégories telles que SN, Sprep, N, V, etc...).

Les correspondances entre suites syntagmatiques, très succintement résumées, sont les suivantes :

Un item t0 correspond à un nom ; un item t1 aux formes nominalisées des verbes : a1 à adjectifs et syntagmes prépositionnels (avec noms ou formes nominalisées comme noyau). Les indicateurs t2 et t3 -et leurs correspondants a2 et a3- incorporent, respectivement, les traits de "machine qui sert à x" et "entité ou individu qui x", x étant le verbe du mot de rappel. Sur le plan syntaxique, t2 et t3 se comportent comme t1 ; a2 et a3 comme a1 ; a4 correspond au participe passé. Les noyaux des syntagmes de phrase correspondent aux noyaux des syntagmes nominaux. Si le noyau est un t1, il correspond aussi aux formes verbales. Le groupe A correspond (1) aux déterminants de ce noyau (V-1 incorpore deux spécificateurs A, l'un qui correspond à un complément de contenu et l'autre de fonction) et (2) au sujet et à tous les compléments de la phrase à l'exception du complément d'objet direct. Le syntagme de phrase qui suit un autre syntagme de phrase correspond au complément

d'objet direct de celui-ci.

Les règles de correspondance sont destinées au documentaliste ; elles constituent un des éléments essentiels du Manuel d'acquisition. Les règles de correspondance syntaxique présentent le schéma général suivant : ce que vous auriez exprimé en LN par la structure x, ou par les structures qui sont en paraphrase avec x, vous devez l'exprimer en V-1 par la structure y. Dans certains cas, lorsque x est ambigu en LN, ce schéma devient : "ce que vous auriez exprimé en LN par la structure x lorsque la structure x possède l'interprétation A, ou par les structures qui sont en paraphrase avec x associée à l'interprétation A, vous devez...". On exige que l'identification des structures LN puisse s'effectuer de manière simple et opérationnelle ; ces structures doivent pouvoir s'induire sans erreur de quelques exemples simples. Tout métalangage compliqué dans la description des suites LN et toute notation complexe dans les suites V-1 sont exclus (en particulier, la notation de parenthétisations hiérarchiques).

On doit donc distinguer, dans le système sous-jacent à V-1, les éléments suivants :

Lexique	Grammaire
Règles grammaticales	
Règles textuelles	
Règles de correspondance	

La naturalité

Les caractéristiques qui définissent la naturalité de V-1 doivent s'expliciter à plusieurs niveaux différents :

(a) les règles de correspondance entre une notion du lexique de V-1 et une invariante de contenu lexématique en LN (cf. mot de rappel en LN).

(b) les règles d'admission d'indicateurs de procédé dans les entrées lexicales de V-1 et la catégorisation syntagmatique du mot de rappel en LN.

(c) les règles de correspondance entre suites syntagmatiques de V-1 et suites syntagmatiques en LN, ces règles de correspondance étant complétées par des jugements de paraphrase et éventuellement d'ambiguïté entre suites en LN.

(d) les emplacements des articulations libres.

(e) les règles générales de remplissage des articulations libres.

(f) les remplissages effectifs des articulations libres par le documentaliste.

Grâce à ces règles de correspondance une suite indiv.$_i$ de V-1 sera associée à un ensemble de suites LN : on le notera F_I^{LN}. On dira que le contenu sign.$_i$ associé à la suite indiv.$_i$ est un quelconque des contenus associés aux suites LN qui appartiennent à F_I^{LN}. En disant "quelconque" on veut indiquer que les distinctions de la LN entre contenus des suites relevant d'une F_I^{LN} sont perdues dans le contenu sign.$_i$.

Les règles de correspondance possèdent donc une double fonction : (1) prescrire au documentaliste les structures de V-1 qu'il doit utiliser pour exprimer un contenu quelconque ; (2) définir les interprétations sémantiques des structures grammaticales schématiques de V-1. Exemple :

Soit la suite indiv.$_i$ de V-1 :

... CATAL ... t1 ... AUTOM ... a1 ... BIBLIOT ... a1

qui correspond, parmi d'autres, à la suite complète :

Catalogage t1 automatique a1 par une bibliothèque a1.

La suite indiv.$_i$ précédente sera associée à une F_I^{LN} qui comporte, parmi d'autres, les suites LN qui suivent ; le contenu sign.$_i$ de la suite indiv.$_i$ précédente sera un quelconque des contenus associés à ces suites:

Le catalogage automatique par les bibliothèques.

Catalogages automatiques dans les bibliothèques.

Catalogage automatique par une bibliothèque.

Aux caractéristiques (a) - (f) ci-dessus qui explicitent la naturalité de V-1 on doit ajouter la suivante :

(g) Par effacement des symboles propres à V-1 dans une suite complète$_i$, on doit obtenir une suite LN appartenant à F_I^{LN}.

Le choix de la naturalité a pour objectif de réduire le coût d'acquisition et d'utilisation de V-1 par le documentaliste et de rendre possible la double utilisation des indexations. Mais la naturalité risque d'introduire dans V-1 des désavantages liés aux phénomènes de paraphrase et d'ambiguïté des LN. On opérera ici au départ avec deux notions très simplistes de ces phénomènes : on considère que l'on a un phénomène de paraphrase lorsqu'un même contenu est associé à deux suites différentes et on a une ambiguïté lorsque deux ou plusieurs contenus (ou interprétations) différents peuvent être associés à une même suite.

Paraphrase

Pour pouvoir adapter la notion de paraphrase à V-1, on doit introduire quelques conventions.

(a) une suite indiv.$_i$ est différente d'une suite indiv.$_j$ ssi les deux suites ne sont pas constituées par les mêmes symboles dans le même ordre. (des entités différentes seront notées par des indices différents).

(b) F^{LN}_I est différente de F^{LN}_J ssi F^{LN}_I est en correspondance avec suite indiv.$_i$ et F^{LN}_J avec suite indiv.$_j$.

(c) F^{LN}_I est distincte de F^{LN}_J ssi il n'y a pas de suite LN appartenant à la fois à F^{LN}_I et à F^{LN}_J.

(d) F^{LN}_I est équivalent à F^{LN}_J ssi toute suite LN qui appartient à F^{LN}_I appartient aussi à F^{LN}_J et vice versa.

(e) F^{LN}_I est partiellement équivalente à F^{LN}_J ssi parmi les suites qui appartiennent à F^{LN}_I et/ou à F^{LN}_J

1) il y en a qui appartiennent à F^{LN}_I et à F^{LN}_J

2) il y en a qui n'appartiennent pas à F^{LN}_I et à F^{LN}_J.

(f) Une suite indiv.$_i$ est en paraphrase totale avec une suite indiv.$_j$ ssi F^{LN}_I est équivalente à F^{LN}_J.

(g) Une suite indiv.$_i$ est en paraphrase partielle avec une suite indiv.$_j$ ssi F^{LN}_I est partiellement équivalente à F^{LN}_J.

Exemple de paraphrase totale (pour faciliter la lecture, les exemples sont présentés sous forme de suites complètes, mais la définition de paraphrase porte sur les suites indiv. correspondantes) :

(1) Ouvrages t0 d'une bibliothèque a1 sur s les mathématiques a1.

(2) Ouvrages t0 sur s les mathématiques a1 d'une bibliothèque a1.

Toutes les suites LN qui, par les règles de correspondance, appartiennent à F^{LN}_1, sont des suites LN qui appartiennent à F^{LN}_2.

Exemple de paraphrase partielle :

(1) Formation t1 pour les bibliothécaires a3.

(2) Formation t1 des bibliothécaires t3.

D'après les règles de correspondance (cf. ci-dessus) on doit exprimer dans le groupe A le sujet et tous les compléments à l'exception du complément d'objet direct. La suite indiv. qui correspond à la suite complète (1)

est par conséquent associée à une F_1^{LN} qui comporte :

Formation par les bibliothécaires.

Formation pour les bibliothécaires.

Mais, dans le schéma général d'une règle de correspondance, on indique non seulement que celle-ci s'établit entre une structure y en V-1 et une structure x en LN mais, aussi, que la correspondance existe entre les structures qui en LN sont en paraphrase avec x, de telle sorte que, dans F_1^{LN}, il faut incorporer, puisqu'en paraphrase avec Formation pour les bibliothécaires, la suite Formation des bibliothécaires. Or, cette suite-ci appartient à F_2^{LN} (Formation t1 des bibliothécaires t3), ce qui n'est pas le cas de Formation par les bibliothécaires.

Les notions de paraphrase totale et partielle peuvent se combiner avec celles de suite syntagmatique et partiellement syntagmatique et avec des quantifications telles que toutes et seulement quelques-unes pour obtenir une typologie des paraphrases possibles en V-1.

Les paraphrases sont fonctionnellement aberrantes : elles ne modifient pas la capacité expressive du langage, car elles re-disent la même chose de manière différente, et elles introduisent du silence, car le système peut être interrogé par une des suites de la famille paraphrastique, alors que l'information demandée est stockée par une autre suite de cette famille. On peut dont se demander pourquoi, puisque la syntaxe est contrôlée, ne pas éliminer du langage le phénomène paraphrastique en modifiant la grammaire. Or, il paraît important de souligner que la solution de l'élimination n'est pas applicable dans la situation actuelle de V-1 si l'on souhaite satisfaire les conditions d'adéquation qu'on s'est données : pour conserver les avantages de la naturalité, et tout particulièrement, si l'on recherche une acquisition rapide et une utilisation afficace par le documentaliste du langage à utiliser, il est malheureusement impossible d'exclure toute forme de paraphrase en V-1.

On discutera d'abord des paraphrases totales. Soit OP_i la famille d'opérations particulières de la grammaire qui conduit à la spécification d'une suite $indiv._i$. Par op (= opération) particulière on comprend tout choix d'un élément et/ou d'une règle de système sous-jacent conduisant à la spécification d'une suite $indiv._i$. Soit la suite $indiv._i$ et la suite $indiv._j$, en paraphrase totale. Puisque ce sont des suites différentes, il faut qu'elles

divergent dans au moins une op. particulière. Pour éliminer la paraphrase entre suite indiv.$_i$ et la suite indiv.$_j$, il faudrait donc éliminer de la grammaire les op. particulières permettant de spécifier comme étant différentes la suite indiv.$_i$ et la suite indiv.$_j$. La grammaire n'engendrerait ainsi qu'une seule suite, ce qui éliminerait du même coup le phénomène de paraphrase. Or, l'impossibilité de ce faire vient des situations rencontrées qui obéissent à un même schéma général, présenté par la suite.

Soient A, B et C, des contenus documentaires possibles et différents. La suite indiv.$_i$ est en paraphrase totale avec la suite indiv.$_j$; elles expriment le contenu documentaire A. Elles sont respectivement spécifiées par OP$_i$ et OP$_j$. Les op. particulières qui apparaissent dans OP$_i$ mais non dans OP$_j$, sont aussi constitutives, avec d'autres, de OP$_k$, qui spécifie la suite indiv.$_k$, celle-ci étant la suite susceptible d'exprimer le contenu B. De même, les op. particulières qui apparaissent dans Op$_j$ mais non dans OP$_i$, sont aussi constitutives, avec d'autres, de OP$_m$, qui spécifie la suite indiv.$_m$, celle-ci pouvant exprimer le contenu C. Il n'existe pas de suites indiv. en relation paraphrastique avec la suite indiv.$_k$ et avec la suite indiv.$_m$, soit parce que ces suites sont impossibles ou extrêmement difficiles à spécifier en V-1, soit parce que la F ^{LN} serait constituée par des suites inacceptables ou à la limite de l'acceptabilité. Cette situation peut se représenter ainsi ("x" indique l'inexistence d'une suite indiv.) :

Grammaire 1

A	B	C
OP$_i$ suite indiv.$_i$	OP$_k$ suite indiv.$_k$	suite indiv.$_n$*
OP$_j$ suite indiv.$_j$	suite indiv.$_l$*	OP$_m$ suite indiv.$_m$

Si l'on modifie la Grammaire 1 en éliminant les op. particulières propres à OP$_i$ (on obtient ainsi la Grammaire 2), on élimine du même coup OP$_k$ et, par là, le contenu B devient impossible à exprimer. Si l'on modifie la Grammaire 1 en éliminant les op. particulières propres à OP$_j$ (on obtient ainsi la Grammaire 3), on élimine du même coup OP$_m$ et c'est C qui devient impossible à exprimer. On a donc :

Grammaire 2

A	B	C
suite indiv.$_i$*	suite indiv.$_k$*	suite indiv.$_n$*
OP$_j$ suite indiv.$_j$	suite indiv.$_l$*	OP$_m$ suite indiv.$_m$

Grammaire 3

A	B	C
OP_i suite indiv.$_i$	OP_k suite indiv.$_k$	suite indiv.$_n$*
suite indiv.$_j$*	suite indiv.$_l$*	suite indiv.$_m$*

Exemple. Les suites (1) et (2) qui suivent sont en paraphrase totale ; elles divergent par l'utilisation du spécificateur A "sur s" en (1) et de "spécialisée a4 b" en (2) :

(1) Bibliothèque t0 sur s la chimie a1.

(2) Bibliothèque t0 spécialisée a4b en chimie a1.

Si l'on supprimait de la grammaire "sur s", il serait impossible d'exprimer dans le langage la distinction entre (3) et (4), et si l'on supprimait "spécialisée a4", on se priverait d'exprimer (5) comme distinct de (6) :

(3) Documents t0 administratifs a1 sur s les bibliothèques a1.

(4) Documents t0 administratifs a1 d'une bibliothèque a1.

(5) Bibliothèque t0.

(6) Bibliothèque t0 spécialisée a4.

On pourrait encore modifier la grammaire de manière à préserver la formulation de B et C, tout en éliminant la paraphrase dans le cas de A. Dans l'illustration précédente, par exemple, il serait nécessaire de formuler les conditions d'utilisation de "sur s" de telle manière que ce spécificateur ne soit pas utilisé entre bibliothèque et chimie. Le grand inconvénient de ce type de solution est la nécessité d'introduire des contraintes particulières et ad hoc dans la formulation d'un op. particulière, qui augmentent le coût d'acquisition et d'utilisation du système et sont sources d'erreur.

Les paraphrases totales ont été admises dans le langage lorsque leur suppression aurait impliqué soit la perte de la capacité expressive du langage (disparition de suites indiv. associées à des F^{LN} distinctes et composées de suites acceptables), soit l'incorporation de règles ad hoc qui alourdiraient le coût d'acquisition et/ou d'utilisation, et à condition qu'une règle paraphrastique soit formulable dans le réseau sémantique de V-1. Une règle paraphrastique ou d'équivalence absolue est une procédure effective permettant de passer automatiquement d'une suite indiv.$_i$ à une suite indiv.$_j$ lorsqu'elles sont en paraphrase totale. Ces règles sont destinées à opérer sur la question documentaire de telle manière que la recherche documentaire s'effectue non seulement en fonction de la formulation primitive de la ques-

tion mais aussi sur les autres formulations possibles de la même question.

On a cherché à formuler des règles paraphrastiques aussi générales que possible, c'est-à-dire portant sur des suites syntagmatiques ou sur des classes de suites syntagmatiques. Celles qu'on commence à formuler portent sur les prédicats symétriques, la linéarité non distinctive introduite dans le noyau à coordination interne et dans le groupe A, les structures engendrées par l'anaphorique, les structures qui incorporent les spécificateurs A et les structures qui incorporent des prédicats. D'autres règles paraphrastiques portent sur des suites partiellement syntagmatiques ou des suites individuelles : elles mettent en général en rapport un item avec un syntagme constitué de deux ou plusieurs items.

Les paraphrases partielles sont en revanche plus gênantes pour le système, puisque des suites LN qui ont des comportements différents sont associées à une même suite indiv. et la machine n'opère que sur celle-ci : il est ainsi impossible de les rattraper par des règles paraphrastiques. Il faut donc éviter ce type d'inconvénient et exclure les paraphrases du langage, sans toutefois supprimer de la grammaire des possibilités distinctives et en essayant de minimiser l'augmentation du coût d'acquisition et d'acquisition.

Il serait extrêmement coûteux d'exclure les paraphrases partielles moyennant des restrictions sur les contextes spécifiques -en particulier lexicaux- où une construction donnée peut être utilisée. Les règles de correspondance deviendraient du type : pour exprimer le contenu x en LN vous devez utiliser la construction y en V-l à condition de ne pas y introduire telle ou telles notions particulières ; si ces notions devaient y apparaître, il faut utiliser la construction z.

On peut obvier à cet inconvénient par l'introduction d'un ordre de priorité dans l'application des règles de correspondance, celles-ci conservant une formulation générale. Par exemple, il est possible de conserver les deux règles générales suivantes :

(1) le syntagme de phrase qui suit à un syntagme de phrase correspond au complément d'objet direct.

(2) tous les autres compléments de la phrase et le sujet s'expriment dans le groupe A

auxquelles on ajoute une instruction explicite dont l'effet est d'obliger le documentaliste à utiliser la correspondance (1) avec une priorité sur la

correspondance (2). De cette manière, si un contenu documentaire peut s'exprimer en LN de manière équivalente comme complément d'objet direct et comme un autre complément quelconque, le documentaliste doit l'exprimer en V-1 au moyen d'une écriture en correspondance avec le complément d'objet direct.

Dans l'état actuel de V-1 on a formulé les règles de priorité suivantes :

(1) Sur les indicateurs de procédé : l'utilisation de t1 et de a1 précède respectivement celle de t3 et de a3.

(2) Sur la structuration de la phrase documentaire : l'utilisation de la correspondance dans chaque ligne précède celle de la ou les lignes suivantes :

2.1 Syntagme de phrase suivant un syntagme de phrase.

2.2 Complément de contenu dans le groupe A.

2.3 Complément de fonction dans le groupe A.

2.4 Noyau à coordination interne.

2.5 Groupe A.

Avec ces règles, si l'on doit contruire en V-1 une expression composée de deux notions, dont la première est un noyau de syntagme et la deuxième aurait pu s'exprimer de plusieurs manières, on est forcé d'introduire dans le système une seule solution. Dans les exemples qui suivent, les suites exclues (= E) ne peuvent être utilisées par le documentaliste comme équivalentes de celles qui sont admises (= A), mais celui-ci est libre de s'en servir pour exprimer d'autres contenus documentaires. Dans la colonne de gauche on a indiqué la priorité qui doit jouer pour obtenir les résultats correspondants (">" = "précédé") :

2.1 > 2.2, 2.5	A : Analyse t1 de la syntaxe t0. E : Analyse t1 portant sur s la syntaxe a1. Analyse t1 syntaxique a1.
2.1 > 2.2	A : Indexation t1 d'un texte t0. E : Indexation t1 portant sur s un texte a1.
2.1 > 2.5	A : Climatisation t1 d'une bibliothèque t0. E : Climatisation t1 dans une bibliothèque a1.
2.1 > 2.5	A : Accroissement t1 du prêt t1. E : Accroissement t1 du prêt a1.
	A : Formation t1 des bibliothécaires.t3. E : Formation t1 pour bibliothécaires a3.

t1 > t3	A : Assistance t1 à la traduction t1.
2.1 > 2.5	E : Assistance t2 au traducteur t3.
	Assistance t1 au traducteur a3.
	Assistance t1 à la traduction a1.

Si le même contenu documentaire ne peut pas être exprimé par a1 et a3, a1 ne précède pas a3 :

	A : Fichier t0 sur s les documentalistes a3.
	A : Fichier t0 sur s la documentation a1.
a1 > a3	A : Manuel t0 sur s l'indexation a1.
2.2 > 2.3, 2.5	E : Manuel t0 pour l'indexation a3.
	Manuel t0 pour s l'indexation a1.
2.2 > 2.5	A : Ouvrage t0 sur s les mathématiques a1.
	E : Ouvrage t0 de mathématiques a1.
2.3 > 2.5	A : Papier t0 pour s la reprographie a1.
	E : Papier t0 reprographique a1.
2.4 > 2.5	A : Bibliothèque t0 et service de reproduction t3 en Belgique a1.
	E : Bibliothèque t0 avec service de reproduction a3 en Belgique a1.

Ambiguïté

Les relations associées à une suite indiv. de V-1 doivent être traitables par la machine et ne peuvent, par définition, être ambiguës. En revanche, une suite LN peut être associée par deux (ou plusieurs) systèmes de relations à deux (ou plusieurs) interprétations différentes. C'est la raison pour laquelle on doit s'autoriser à formuler les règles de correspondance entre constructions de V-1 et constructions en LN en prévoyant les ambiguïtés de celles-ci. Par exemple, si un document parle de

Bibliothèques universitaires et municipales

il peut faire référence soit (1) à des bibliothèques qui sont en même temps universitaires et municipales, soit (2) à certaines bibliothèques qui sont universitaires et d'autres bibliothèques qui sont municipales. Pour éviter cette ambiguïté en V-1, l'instruction qui spécifie l'utilisation des items A dans le groupe A exige que tous les items A doivent porter sur une même entité référentielle, autrement dit, qu'on doit seulement exprimer dans un même

groupe A l'interprétation (1) ci-dessus.

Sans essayer d'épuiser la complexité de la relation entre l'ambiguïté en LN et ses conséquences en V-1, on attirera l'attention sur deux points.

1er point. En accord avec le sens général de l'analyse de l'ambiguïté par Pierre Le Goffic [1], les règles de correspondance entre suites V-1 et suites LN, lorsque celles-ci présentent un phénomène d'ambiguïté, ne peuvent pas s'exprimer par des instructions simples du type "si l'interprétation est A vous devez utiliser la construction x et si l'interprétation est B, vous devez utiliser la construction y".

En effet, on doit prévoir au moins les situations suivantes :

Soit A et B deux interprétations distinctes.

Soit les notions de suite syntagmatique LN et suite individuelle LN (intégrée celle-ci par des lexèmes particuliers).

(1) Dans tous les cas, à côté de A et B, il faut prévoir l'interprétation "je ne sais pas" (notée "?").

(2) Dans certains cas, on a une des deux situations suivantes :

(2-1) Certaines suites individuelles relevant d'une suite syntagmatique déterminée présentent une interprétation A ou bien B, alors que d'autres présentent une valeur C mixte, fonction de A et de B. Par exemple, le cas de

Coopération nationale et internationale d'une bibliothèque

ne peut être traité comme l'exemple précédent (il paraît en effet assez vain de se demander s'il s'agit d'une même coopération ou de deux coopérations distinctes sur le plan référentiel) mais doit plutôt être interprété par un contenu du type : coopération unique (puisque d'une seule et même bibliothèque) mais qui compte deux aspects différents.

(2-2) Les suites individuelles relevant d'une suite syntagmatique déterminée présentent dans un contexte une interprétation A ou bien B, alors que dans d'autres contextes elles présentent une interprétation A et B. Par exemple, la traduction.

Pour résoudre les problèmes que les cas précédents soulèvent, on devrait introduire en V-1 les distinctions suivantes :

---

1 Ambiguïté linguistique et activité de langage. Thèse. Université de Paris VII. Paris, 1981.

suite indiv.$_i$: A
suite indiv.$_j$: B
suite indiv.$_k$: "?"
suite indiv.$_l$: A et B
suite indiv.$_m$: C

Or, d'une part, ce serait très coûteux sur le plan de l'acquisition et de l'utilisation du langage, et d'autre part, les exigences de la naturalité ne seraient pas satisfaites : il n'y a pas toujours en LN des structures simples et distinctes pour exprimer les cinq solutions possibles, les interprétations respectives en LN résultant d'une interaction complexe entre les contenus lexématiques et les structures syntaxiques. On a donc incorporé à V-1 les solutions suivantes :

(a) suite indiv.$_i$: A et B, A et (? B)
suite indiv.$_j$: A
suite indiv.$_k$: B
(traitement de l'indication de procédé t1)

(b) suite indiv.$_i$: A
suite indiv.$_j$: B, ?, C
(traitement du noyau à coordination interne, des constructions qui portent sur lui et du groupe A en relation avec le noyau de syntagme de phrase).

Le fait que l'utilisation de V-1 soit redevable de la connaissance des règles de correspondance, présente des avantages pour l'analyse documentaire : en guidant le documentaliste dans son travail d'écriture, elles l'obligent en même temps à approfondir sa compréhension du texte traité. Mais cette compréhension trouve ses limites dans le "flou" du texte lui-même, qui devient ainsi incorporé au langage documentaire. Et on retrouve ainsi, dans l'outil même de documentation, une autre source de perturbation du système, que l'on peut certes essayer de réduire mais que, dans le cadre des contraintes qu'on s'est données, il semble impossible d'éliminer.

2ème point. Il existe des cas particuliers en LN qui relèvent de la notion générale de paraphrase, car une même interprétation peut être associée à deux systèmes distincts de relation, mais qui relèvent aussi de la notion générale d'ambiguïté, car une même suite peut être associée à ces deux systèmes différents de relation. Exemples :

(1) L'accroissement du prêt

(2) Salle de lecture d'une bibliothèque en Allemagne.

Dans (1), prêt admet la relation de sujet et objet d'accroître sans modification de l'interprétation ; de même, en Allemagne en (2) peut porter sur salle de lecture et sur bibliothèque, avec une même interprétation. (pour des exemples analogues, cf. Le Goffic, op. cit., p. 185, et leur caractérisation en termes de neutralisation, p. 382).

On a la situation générale suivante, caractérisée par :

(a) une suite syntagmatique déterminée.

(b) deux systèmes de relation différents opérant sur cette suite

(c) certaines suites individuelles qui relèvent de la suite syntagmatique sont associées par chacun de ces systèmes à une interprétation différente (ce sont les cas classiques d'ambiguïté : la cruauté de l'ennemi, succursale d'une entreprise en Allemagne), alors que d'autres suites individuelles relèvent de la même suite syntagmatique sont associées par chacun de ces mêmes systèmes de relation à une même interprétation, ce sont les cas mixtes des exemples (1) et (2) ci-dessus, relevant à la fois de la paraphrase et de l'ambiguïté.

Si le phénomène apparaît en LN dans une suite syntagmatique composée de deux entités telles que chacune correspond à un item ou à un syntagme différent en V-1, et si les relations dans la suite syntagmatique en LN peuvent s'exprimer par deux suites individuelles différentes en V-1, les règles ordonnées de correspondance présentées ci-dessus permettent de résoudre le problème ; c'est le cas de l'exemple (1) qui doit s'exprimer en V-1 par (1-1) et non par (1-2), ce qui règle la question vue sous l'angle de l'ambiguïté, tout en évitant l'introduction de paraphrase :

(1-1) L'accroissement t1 du prêt t1.

(1-2) L'accroissement t1 du prêt a1.

En revanche, si le phénomène apparait en LN dans une suite syntagmatique composée de trois entités (ou plus), le problème est autrement difficile à résoudre. En effet, si l'on a une suite x y z, celle-ci admet les systèmes de relation (1) et (2), associés à un même contenu :

(1) [x [y z]]

(2) [x [y] [z]]

Si l'on admet en V-1 les écritures qui correspondent à (1) et à (2), on

introduit une paraphrase totale. Or, dans l'état actuel de ce qu'on connait, ce type de paraphrase porte sur des suites individuelles et non sur des suites syntagmatiques. La paraphrase est liée au contenu des entités x, y et z ; elle existe dans (1) mais non dans (2) ci-dessous, et le calcul a priori des cas paraphrastiques ne semble pas simple :

(1) Ecrire un roman en russe
(2) Traduire un roman en russe
x y z

Si par le biais de restrictions ad hoc dans les règles de correspondance on élimine une parmi les deux écritures possibles -c'est-à-dire, si l'on envisage une solution du type de celle évoquée ci-dessus, introduisant un ordre dans les correspondances- on élimine la paraphrase mais on n'introduit pas moins un facteur de silence. En effet, pour une interrogation partielle, le document est pertinent et pour [x z] -écrire en russe- et pour [y z] -roman en russe- : avec une seule écriture possible on élimine une de ces deux possibilités.

Si, par le biais d'instructions ad hoc dans les règles de correspondance, on demande au documentaliste d'écrire [x z] et [y z], le résultat est non seulement l'augmentation du coût d'acquisition et d'utilisation mais, aussi, la non satisfaction d'une des exigences de la naturalité : la suite complète correspondante (Ecrire t1 en russe a1 un roman t0 en russe a1) ne peut pas être associée par effacement des symboles V-1 à une suite acceptable en LN.

Dans un sondage en échelle réduite[1] on a essayé de tester si des documentalistes différents, opérant avec V-1 sur les mêmes documents, et après avoir fixé l'interprétation de ces documents, proposaient une même indexation. Trois documentalistes ont ainsi produit 69 indexations à partir de 23 résumés en LN. On a considéré que le 100 % de coïncidence était atteint si les 3 documentalistes indexaient de manière identique l'ensemble du document. Sur l'ensemble des textes, on a obtenu une coïncidence absolue de 82 % ; ce sont des indexations où soit l'écriture primitive en V-1 est identique, soit des règles paraphrastiques permettent d'associer des écritures différentes. A cette coïncidence absolue s'ajoute un 9 % de coïncidence approximée : les règles

---

1 Effectué en collaboration avec F. LAGUEUNIERE et J. SAUZEDDE.

qui les produisent, à l'instar de règles paraphrastiques, opèrent sur des écritures différentes, mais mettent en relation des structures véhiculant une information très proche mais non identique.

Les résultats de ce sondage -pratiqué certes en échelle très réduite et qui demandent à être confirmés- montrent qu'il y a des raisons de penser qu'il devrait être possible d'éviter avec V-1 certains obstacles qui empêchent l'uniformité des indexations par des documentalistes différents. Mais il est significatif que tous les cas de non coïncidence (environ 10 %) relèvent du point (2) évoqué ci-dessus à propos de la paraphrase. Cf. par exemple :

Programme d'enseignement pour bibliothécaires.
Utilisateurs de l'information dans les entreprises métallurgiques.
Réseau national de documentation en Irlande.
Fonds des archives de littérature en Allemagne.

Ce problème est sans doute en rapport avec la question plus générale des interrogations partielles -vraisemblablement les plus fréquentes- qui portent sur deux ou plusieurs entités dans une suite qui en comporte d'autres. Pour résoudre ce problème il faudrait mieux comprendre les questions de "portée" entre entités d'une suite de surface LN.

Sur le plan général, on peut se demander si en incorporant aux langages de représentation du contenu des documents un certain type de relations syntaxiques propres à LN et n'incorporant que ces relations, on ne s'approche pas d'une situation limite, où on pourrait montrer que, quoiqu'on en fasse, le bruit et le silence ne diminueront pas au-delà d'une certaine proportion dans la recherche documentaire. En continuité avec l'analyse du problème par Jacques Maniez[1], on doit alors se demander si la réduction ainsi obtenue de bruit et/ou de silence mérite l'introduction de ces relations. Il semble bien qu'ici comme partout ailleurs dans le traitement des problèmes du langage, on est en présence d'un problème de "coûts conflictuels". Il paraît aussi clair que ce problème général et les problèmes plus spécifiques avant évoqués sur la paraphrase et l'ambiguïté des LN, vont se poser à toute recherche documentaire et à toute représentation du contenu des documents, qu'elles soient faites "manuellement" ou par la machine.

Université de Clermont II
C.I.L.N.

1 Le rôle de la syntaxe dans les systèmes de recherche documentaire. Thèse. I.U.T. de Dijon,

DISCUSSION

T. ANGELETTI

Je reviens sur "écrire un roman en russe / traduire un roman en russe". Il me semble qu'on pourrait s'en sortir par des considérations sur lexique/grammaire (cf. la communication de J.P. Desclés). En effet, il se pose un problème de modalité au sens large, c'est-à-dire portant sur le rapport d'un objet métalinguistique à un repère énonciatif. Avec "écrire", on a un processus qui marque le passage d'un état où le roman n'existait pas à un état où il existe comme "(étant) écrit" : la notion "roman" est reliée d'une façon particulière à une situation énonciative qui représente l'état "avant le début de la traduction". On peut bien sûr se demander quelle propriété il avait alors : russe, allemand, italien..., et s'il avait la propriété "russe", quelle propriété il a après la réalisation du processus...

Bref, on voit que le lexical et le grammatical sont interdépendants. Et que le problème est bien d'automatiser ces gloses intuitives.

M. FISCHER

1. Sous une forme un peu brutale, j'ai l'impression que le but de V-I est de faire travailler l'homme pour la machine. Dans ces conditions, quelle en est l'acceptance auprès des utilisateurs potentiels ? J'avoue que comme ancien documentaliste j'aurais beaucoup de réticences à truffer ainsi le LN par des marqueurs destinés à la machine. J'accepterais mieux de renoncer à certaines formulations génératrices de paraphrases (par exemple, actif à la place de passif) sans, pour autant, tomber dans le corset des 17 formes normales de TITUS.

2. A moins d'avoir mal entendu, j'ai l'impression que V-I ne s'applique qu'à la langue française. Il ne me semble pas prendre en compte la juxtaposition des substantifs (sans aucune indication du cas grammatical ni même de liens ambigus via des prépositions) comme on les trouve en anglais -et encore moins des mots composés en allemand- livrés à la créativité la plus débridée.

3. En référence à ce qui précède, pourquoi s'acharner à lever de manière automatisée des ambiguïtés qui sont souvent inhérentes à la LN ou du moins

au texte particulier qui ne comporte pas toujours d'information pour pouvoir les lever. Qui n'a pas relu 3 fois une expression composée anglaise ou un mot composé en allemand, avant de penser avoir appréhendé la bonne signification ?

Gabriel G. BES

L'ordre des questions a été modifié.

V-1 ne s'applique, dans son stade actuel, qu'à la langue française (cf. point 2 de M. Fischer). Les relations créées par les structures choisies devraient cependant être suffisamment générales pour les retrouver dans d'autres langues, et notamment en anglais et en allemand. Mais on n'a pas encore étudié quelle devrait être l'écriture par le documentaliste de ces structures dans les langues mentionnées.

Je ne pense pas qu'on s' "acharne à lever de manière automatisée des ambiguïtés" (cf. point 3 de M. Fischer). L'exposé va plutôt dans le sens contraire : les ambiguïtés doivent être reconnues par le documentaliste ; on lui demande d'utiliser des structures différentes selon les interprétations, et on prévoit justement le cas où l'ambiguïté est impossible à lever (cf. la notation "?").

Est-ce qu'il serait utile de laisser s'introduire dans le système davantage d'ambiguïté, c'est-à-dire, de ne pas demander au documentaliste d'en lever quelques-unes ? Je renvoie, pour cette question, au point (c) des conditions d'adéquation et à la fin de l'exposé : il ne me paraît pas utile de se poser cette question dans un cadre général, a priori valable pour toute situation : plus fines seront les discriminations syntaxiques, et plus rigoureuses et compliquées les instructions aux documentalistes, moins se feront sentir les effets documentaires néfastes de l'ambiguïté et de la paraphrase... mais le système deviendra plus coûteux : en temps d'acquisition par le documentaliste, dans la définition des règles effectives qui doivent constituer les familles paraphrastiques, dans son informatisation. Le "1" de V-1 indique un état d'équilibre, proposé comme hypothèse de travail et qui doit être adapté aux exigences -et aux moyens- d'un éventuel centre de documentation.

Parfaitement d'accord avec M. Angeletti sur le fait que "le lexical et le grammatical sont interdépendants". Mais, une fois acquit ce point très général, je deviens beaucoup plus sceptique sur les possibilités de "s'en sortir par des considérations sur lexique/grammaire". Ces "considérations"

ne me semblent pas évidentes, et, par ailleurs, je ne les trouve pas dans la riche communication de Desclés (cf. mes commentaires p. 103) autrement que sous forme très programmatique. Restons dans le cadre de l'exemple commenté et adoptons le système d'analyse esquissé pas Angeletti. Lire, citer et transcrire, tout comme "traduire", exigent l'existence du complément d'objet direct avant le commencement du processus ; cependant, seulement citer et transcrire se comportent comme traduire alors que lire est du type écrire. Le comportement d'écrire n'est d'ailleurs pas toujours le même vis-à-vis du problème évoqué, même si son objet direct reste le fruit de l'écriture : écrire sans commentaire un rapport et écrire un rapport sans commentaire, écrire sans difficulté excessive un manuel et écrire un manuel sans difficulté excessive, ne sont pas dans la même relation qu'écrire en russe un roman et écrire un roman en russe. Evoquons seulement les compléments adverviaux qui modifient le sens d'écrire : dans écrire sous la dictée un roman en russe, il n'est plus forcément vrai que le roman n'existait pas avant le processus d'écriture, mais écrire sous la dictée conserve, par rapport à un roman en russe, les mêmes propriétés paraphrastiques que écrire. Encore une autre difficulté : le facteur de l' "existence"-maintenant entre guillemets- du référent associé au complément d'objet direct à un moment donné et par rapport à un repère énonciatif, doit très probablement être un des facteurs, qui s'intègre à bien d'autres, et dont on devra tenir compte dans une solution du problème. Mais quelle est la capacité opérative de cette notion ? Que faire par exemple avec il a écrit pour la troisième fois le dernier chapitre de sa thèse ou avec il a écrit aujourd'hui la même solution que toi hier ? Le dernier chapitre de la thèse et la solution "existaient"-ils avant de les écrire ? Ceci nous amène au premier point soulevé par M. Fischer.

Contrairement à son impression, l'objectif de V-1 n'est pas de "faire travailler l'homme pour la machine", mais, plutôt, d'essayer de trouver l'équilibre le plus efficace possible entre ce qu'on peut demander à l'un et à l'autre de faire dans l'état actuel des connaissances. Qu'impliquent sur le plan documentaire et sur celui de l'analyse automatique les problèmes évoqués ?

Soit la question des paraphrases entre deux entités, pour laquelle on a proposé une solution fondée sur des règles ordonnées de correspondance. Certes, c'est malheureusement une solution qui, sur le marché actuel du traitement automatique, serait considérée comme fort peu élégante. Que faire pour l'automatiser ?

Le lexique de V-1 est réduit : environ 1 000 notions (calcul bas, puisque chaque notion admet plusieurs indicateurs de procédé). L'ordre étant pertinent, on a 1 000 000 de groupes binaires différents qui sont en principe possibles. Il y a 5 relations syntaxiques différentes (mais dans l'une l'ordre n'est pas pertinent). Chacun des groupes binaires constitué avec une des relations syntaxiques doit être comparé au même groupe avec les 4 restantes. A la suite de plusieurs millions de comparaisons (moins de 10, car tous les groupes n'admettent pas toutes les relations syntaxiques), on aura découvert les paires paraphrastiques.

Soit le problème de la paraphrase-ambiguïté entre trois notions. Ce problème n'apparaît pas dans les 5 relations syntaxiques possibles dans les groupes binaires ; on l'a détecté dans seulement deux, mais, pour simplifier, on ne considère que le million de groupes binaires en principe possibles : pour découvrir les cas où le phénomène apparaît et les dissocier des autres, on doit traiter 1 milliard de groupes.

Il est clair que même dans un langage aussi rudimentaire que V-1, la question ne peut être résolue par des tables ou des descriptions a priori des cas particuliers. On ne peut pas visualiser le problème en imaginant qu'il s'agit, avec un peu de patience, de remplir 1 000 entrées d'un lexique. Et on ne connait aujourd'hui aucune théorie générale suffisamment riche, avec des contraintes explicites, permettant de calculer a priori, un lexique et une syntaxe étant définis, et un domaine d'application non artificiellement local étant donné, quels seront les cas paraphrastiques et/ou ambigus pour les distinguer des autres. Certes, il y a des notations des items et des structures individuelles qui devraient permettre de se donner des règles de passage entre structures avec des notations différentes. Mais, autant qu'on sache, il n'y a aucune possibilité de prévoir les cas paraphrastiques ou d'ambiguïté dans des textes qui ont une fonction pratique quelconque, tant soit peu élargie et non artificiellement délimitée. Et, sur le plan heuristique, on peut se poser la question de l'opportunité des outils de notation qui sont proposés, alors que l'on ne dispose pas des moyens adéquats pour constituer un corpus d'observations qui vont les tester et, a fortiori, sans qu'on ait une évaluation plus précise du problème dans un secteur non artificiellement délimité.

Autant qu'on sache, il n'existe pas aujourd'hui des inventaires complets, exhaustifs, de toutes les situations d'ambiguïté et de paraphrase dans un cadre syntaxique qui soit un peu large et intéressant. En ce qui nous concerne,

nous arrivons mal à faire le tour complet des problèmes qui se posent dans le SN, tel qu'il se répercute en V-I. Par ailleurs, autant qu'on sache, on n'a aucun outil sérieux pour mesurer avec un peu de précision la répercussion dans un domaine technologique donné -par exemple, dans la recherche documentaire- des phénomènes étudiés. Un bon indice que ces problèmes existent -et qu'ils gênent- est le petit sondage réalisé : tous les cas détectés de non uniformité d'indexation viennent de là. La plupart des exemples que cette discussion a provoqué ne sont pas des inventions de laboratoire du linguiste : il suffit de consulter le Thesaurus et le Bulletin signalétique des Sciences de l'Information du CNRS pour les trouver abondamment.

Si l'on admet cette situation, on peut se poser la question de savoir si une tentative d'analyse entièrement automatique est tellement plus "scientifique", "généralisante" et "intéressante". Restons dans notre sphère d'application. Admettons qu'une analyse automatique réussisse l'analyse morphologique et l'analyse syntaxique des textes traités. Admettons qu'on ait ainsi en sortie des arbres avec des noeuds étiquetés ou des structures équivalentes (c'est loin d'être aujourd'hui le cas). Que faire avec les familles paraphrastiques dans ces structures ? Si on les ignore, on introduit systématiquement le silence. Si les familles paraphrastiques sont constituées d'une manière grossière (par exemple, si les notions A et B sont demandées dans la relation x, la machine pourrait construire toute la famille de structures où les notions A et B apparaissent dans toutes les autres relations syntaxiques possibles), on introduit le bruit : on de-construit, en quelque sorte, le résultat de l'analyse syntaxique.

Certes, on peut justifier l'analyse automatique en tant qu'heuristique permettant de trouver dans les textes les problèmes intéressants. On se donne un algorithme et, dans d'autres occasions que dans les séances de démonstration grand public, on l'utilise pour tester les cas où il ne marche pas. Si on les considère assez "fréquents" ou "normaux", on se décide à améliorer l'algorithme, etc. On peut se demander, sur le plan de la gestion de crédits, si c'est une méthode tellement rationnelle : est-ce qu'une analyse qualitative a priori plus un programme de simulation simplifié ne donneraient pas les mêmes résultats et de manière beaucoup plus économique ? Il est regrettable que ce type de problème ne soit presque jamais abordé ouvertement, et on peut le laisser maintenant de côté. Sur un autre plan, on peut se demander que signifie une statistique portant, d'une manière ou d'une autre, sur des phénomènes syntaxiques. Et il reste encore un problème qui me semble être le

plus délicat. Est-ce que par un procédé inductif -"en regardant beaucoup de texte" ou "en regardant toujours un peu plus de texte"- on va dégager les observations qui doivent conduire à la présentation de ce puzzle gigantesque qu'est un système linguistique? En tout état de cause, je ne crois pas que ce soit en ajoutant des "détails" à des univers très restreints où l'on fait marcher un programme, que l'on s'approchera d'une solution applicable à un problème en grandeur nature.

Dans la préparation de V-1 on a très consciemment suivi une heuristique contraire à celle qui semble aujourd'hui dominante. Avec lui on peut, en effet, simplifier et circonscrire des problèmes de traitement de l'information d'un texte grâce notamment à la contrainte de naturalité, sans pourtant dénaturer le domaine d'application, qui reste de grandeur réelle. Pour traiter ces problèmes, on n'a pas choisi une tactique de tout ou rien (ou le documentaliste ou la machine) mais on cherche à se donner des critères permettant d'aboutir au point d'équilibre le plus efficace dans une situation donnée. Connaissant les structures syntaxiques qu'on se donne, on pense pouvoir calculer les problèmes paraphrastiques généraux qui peuvent se poser. On devrait ainsi pouvoir commencer à simuler et à mesurer les répercussions documentaires de ces problèmes. D'autre part, on devrait pouvoir aussi envisager les solutions possibles (y compris l'absence de solution), et leur coût et conséquences. Par exemple, il est clair que l'information qu'on incorpore dans les entrées lexicales -que ce soit dans une base de connaissances ou dans la grammaire- va jouer un rôle dans la constitution des paraphrases. Quelle complexité dans la description de ces entrées est requise pour résoudre quelles paraphrases et avec quelles répercussions sur le système documentaire? Voilà la question qui me paraît essentielle. Si on trouve des solutions entièrement traitables en machine, elles seront évidemment les bienvenues : on pourra décharger d'autant le documentaliste. On déplacera ainsi le point d'équilibre, diminuant son travail. Si on ne les trouve pas et que le documentaliste considère qu'on lui laisse une tâche trop lourde, le système sera défectueux et ne remplira pas ses fonctions dans les limites assignées (cas évoqué par M. Fischer dans son point 1). Mais on aura connu de manière analytique l'origine d'un problème. Et c'est un facteur qui me paraît important dans la justification de l'heuristique de V-1 : s'il passe les tests d'adéquation -parmi lesquels celui de l'acceptation pour le documentaliste (cf. le point 1 de M. Fischer)- il devrait être utile sur le plan documentaire.

Mais si, sur ce plan, échecs il y avait -et échecs au pluriel, caractérisés le plus analytiquement possible- V-I devrait servir pour caractériser les origines de ces échecs. Et les réussites et les échecs devraient être généralisables, ce qui paraît important.