POLYSEMIE ET POLYMORPHIE

Delatte

Nous avons dit hier à propos de la polysémie et de la polymorphie tout ce que nous avions à dire. Nous avions cru, assez naïvement, qu'en faisant une analyse contextuelle, nous pourrions arriver à un résultat, mais les quelques expériences que nous avons faites montrent qu'on ne peut pas aller très loin. Nous en avons donné toute une série d'exemples, aussi bien dans le domaine de la syntaxe que dans le domaine de la morphologie.

Tous les essais montrent qu'il est impossible, par une étude du contexte, de supprimer les amphibologies avec un programme. Il faudrait faire intervenir le sens, la sémantique. Si on pouvait le faire, ce serait de la traduction automatique. La seule chose possible, c'est de faire intervenir les probabilites. On dirait que telle analyse a 90°/$_0$ de chances d'être correcte. Mais nous faisons de la philologie et non pas de la traduction automatique. Or, en philologie, il ne m'intéresse pas de savoir que *rosae* est le génitif singulier avec une certaine probabilité, je dois le savoir avec certitude.

La philologie, c'est précisément ce qui n'est pas soumis aux probabilités car c'est lá que se trouvent les effets de style et peut-être les critères qui sont nécessaires pour traiter les problèmes purement philologiques de chronologie relative, d'authenticité et des mécanismes littéraires.

Voilà ce que nous pensons, mais peut-être avez-vous sur cette question des idées meilleures que les nôtres.

Busa

Vous avez donné des faits qui montrent qu'il n'est pas possible de résoudre des amphibologies en analysant le contexte par programme, n'est-ce pas? A propos de la polysémie, il y a aussi d'autres points à exposer. Une première question: quand s'agit-il de la polysémie, qui est de la véritable homographie, au sens étroit, et quand s'agit-il de la polysémie qui n'est pas de l'homographie?

Cette question pourrait être traduite aussi comme ça: quand un lemme reste-t-il un lemme avec plusieurs sens différents et quand s'agit-il au contrai-

re de plusieurs lemmes, chacun avec un sens différent? A ce propos à Gallarate nous n'avons rien décidé, sauf d'accepter la division des lemmes proposée dans le Forcellini. En second lieu il y a les catégories ou niveaux d'homographie. Pour nous, il y en a quatre qui ont résisté à l'élaboration électronique, et je les répète:

a) homographie entre mots de différentes langues;
b) homographie entre lemmes dont la coïncidence est accidentelle, due au hasard;
c) homographie entre lemmes qui ont une signification radicale commune, mais qui appartiennent à des paradigmes différents: c.à.d. dont le sens commun de fond est différencié morphologiquement;
d) homographie entre formes à l'intérieur d'un même paradigme. En connexion avec ce point, on a envisagé une recherche sur l'origine de l'homographie: des homographies sont créée par les désinences, d'autres par les préfixes, et enfin d'autres sont fondées sur la partie invariable du mot.

Il y a des cas spéciaux d'homographie auxquels il faut faire attention: avant tout l'onomastique; puis l'homographie des enclitiques; troisièmement: les formules. Dans les textes de S.Thomas, qui sont des commentaires de logique ou de géométrie, on parle par ex. de la ligne AB: dans le texte, AB est imprimée en lettres majuscules, mais elle est perforée comme la préposition *ab*. On parle d'un triangle PRO: on a des milliers de cas comme ça. A des situations semblables nous avons appliqué pendant la rédaction, un code: tout ce qui est formule a été cerclé en rouge et perforé avec un code spécial. Autrement on devait trier à la main tous le I, qui sont l'impératif du verbe *eo, ire* et ceux qui sont le numéral, tous les II qui sont un numéral et ceux qui sont le pronom etc.

Un dernier point à discuter sur la polysémie, c'est la méthode pour la découvrir. A la limite, c'est clair, on devrait faire la flexion automatique de tous les lemmes qui existent par exemple dans le Forcellini, et après en sélectionner alphabétiquement toutes les formes; on trouverait toutes les formes qui sont des homographes possibles. Mais on en sortirait beaucoup plus de 15 millions des formes!

Selon ma liste, la première conjufaison régulière, y compris les participes et les comparatifs et superlatifs des participes, a 431 flexions possibles, la deuxième conjugaison 480, la troisième 408, et la quatrième 448.

Pour les déclinaisons régulières, comparatifs et superlatifs inclus, ainsi que les terminaisons du nominatif, elles sont 50 pour la première, 165 pour

la deuxième, 294 pour la troisième, 156 pour la quatrième, 127 pour la cinquième.

Supposons alors que nous ayons une moyenne de 200 flexions pour chaque lemme. Les lemmes du Forcellini sont 90.000!

Delatte

Oui, vous pensez à retrouver les formes amphibologiques en développant la flexion complète de tous les mots latins contenus dans Forcellini. Chez nous, cela se fait automatiquement. Il y a parmi les documents que nous avons apportés un listing d'analyse automatique. Il serait intéressant de l'examiner. Vous verrez comment l'ordinateur révèle les amphibologies, beaucoup mieux que nous ne pourrions jamais le faire, parce que sa logique implacable le conduit à tout voir, y compris les choses qu'on ne voit pas d'habitude. Il s'agit de l'analyse d'un passage de Lucrèce.

Evrard

Dans les documents du LASLA, vous trouverez, après les échantillons des deux lexiques et des tables de désinences, un document qui s'intitule "Listing d'analyse automatique". Nous allons voir comment l'ordinateur a procédé, par exemple, pour le troisième mot, *"summa"*. Si on demandait à un latiniste quelles sont les analyses possibles de "summa", je doute qu'il en trouve douze. D'ailleurs, les douze analyses que vous voyez sur le listing ne sont pas enregistrées comme telles dans la mémoire de l'ordinateur. En fait, on n'y trouve que les éléments (radicaux et désinences), et c'est la tâche propre de notre programme d'analyse de combiner ces éléments. Tout d'abord, le lexique enregistré sur disque contient le substantif *summa* et précise que ce mot appartient à la 1ère déclinaison et a pour radical *summ*. Par ailleurs, le programme, lorsqu'il trouve une forme identique au lemme d'un mot de la 1ère déclinaison, sait qu'il peut l'analyser comme nom. ou voc. sing. de ce mot. Enfin, la table des désinences contient une désinence *-a*, qu'elle définit comme étant une désinence d'abl. sing. pour les mots de la 1ère déclinaison. C'est en combinant tous ces renseignements que l'ordinateur fournit les trois premières analyses, dont les codes 11F, 11A et 11B signifient respectivement

abl. sing., nom. sing. et voc. sing. d'un substantif de la lère décl. Voilá donc trois analyses et trois lemmatisations possibles de *"summa"* que l'ordinateur fournit alors que nous ne les avons pas mises comme telles dans l'ordinateur.

Le lexique enregistré contient ensuite un substantif *summum*, radical *summ*, neutre de la 2e décl. Comme la table des désinences offre une désinence *-a*, caractéristique de l'acc. plur. des neutres de la 2e décl. l'ordinateur, en combinant ces deux données, a formé la 4e analyse, codée 12L, ce qui signifie substantif de la 2e décl. à l'acc. plur. Enfin, l'ordinateur connaît par ses mémoires un superlatif irrégulier de *superus* qui a pour radical *summ*. Comme par ailleurs, la table de désinences contient une désinence *-a* caractéristique du nom. sing. féminin du superl., il peut former la 5e analyse, dont le code (2JA002) signifie qu'il s'agit d'un superlatif dont le positif appartient à la lère classe, au nom. sing. fém. Les analyses qui suivent réutilisent deux lemmes déjà mentionnés, *summum* et *superus*, mais avec d'autres désinences.

Il n'est sans doute pas nécessaire que j'en fasse l'énumération. En résumé, il nous a suffi de mettre dans les mémoires un adjectif *superus* ayant comme thème *super* et comme superlatif *summ*, un substantif de la première déclinaison *summa* et un neutre de la deuxième déclinaison *summum* et d'autre part d'avoir enregistré les différentes désinences.

C'est l'ordinateur qui réalise la jonction des éléments qui sont compatibles. Ainsi, toutes les formes de tous les radicaux enregistrés dans le lexique sont analysées virtuellement. Dès lors, toutes les homographies possibles sont découvertes par l'ordinateur, dans la mesure où les lexiques et la table des désinences sont complets.

Sans doute peut-il arriver qu'une lacune de l'un des lexiques empêche de découvrir une homographie. Mais le cas est rare, puisque les mots omis dans nos lexiques sont des mots très peu employés. De plus, le contrôle philologique des listings est précisément destiné à découvrir des cas de ce genre.

Dans cette prospective, les niveaux d'homographie n'ont pas pour nous d'intérêt pratique. Quelle que soit la raison pour laquelle il y a l'homographie, du moment qu'il y en a une, il faut que nos lexiques de mots et de formes et notre table de désinences, soient constitués de manière à pouvoir la découvrir. Qu'il s'agisse d'une homographie à l'intérieur d'un même lemme

par le jeu des désinences, ou de l'homographie qui résulte d'une coïncidence de formes venant de radicaux différents, la chose n'a pas d'importance pratique pour nous.

Busa

Alors vous êtes d'accord que pour découvrir le cadre complet de l'homographie il faudrait avoir recours à la flexion automatique des lemmes latins, n'est-ce pas? L'étude des niveaux de l'homographie, ne sert pas pour la pratique de l'analyse; mais est une recherche pure sur le système de la langue latine.

Un autre point à ajouter est la distinction entre les degrés de probabilité de vérification des homographes.

Evrard

A propos des degrés de probabilité, il nous paraît impossible de nous y fier entièrement. Prenons par exemple le cas de *"est"*, qui est habituellement une forme de *"sum"* et, de loin en loin, une forme de *"edere"*. Ce qui nous intéresse, ce sont précisément les apparitions de cette forme tout à fait exceptionnelle. Cependant nous tenons compte des probabilités de la manière que voici. Dans le listing d'analyse automatique que j'ai commenté tantôt, certains de vous se sont peut-être demandé pour quelle raison les analyses se succédaient dans un ordre à première vue incompréhensible. C'est que, dans la mesure du possible, nous nous sommes efforcés de faire en sorte que les analyses apparaissent en ordre de fréquence. Ainsi, nous savons que l'ablatif est nettement plus fréquent que le nominatif, et que le nominatif à son tour est beaucoup plus fréquent que le vocatif. Dès lors, la première analyse est celle de l'ablatif, puis viennent celle du nominatif et celle du vocatif.

L'avantage, c'est que, lorsqu'un philologue reçoit un listing analogue à celui dont vous avez une photocopie, et qu'il doit choisir, parmi toutes les propositions de l'ordinateur, celle qui est correcte, on peut dire que dans huit ou neuf cas sur dix, la bonne solution se trouve à la première ligne, ou en tout cas à l'une des deux premières lignes. A ce propos, permettez-moi de revenir sur ce que M. Perschke disait hier relativement à l'analyse en vue de la traduction automatique.

Vous disiez que dans la décomposition d'une forme en radical et désinence, vous prenez comme principe d'adopter la division qui donne le radical le plus long et la désinence la plus brève.

Perschke

Oui.

Evrard

Je comprends bien pour quelle raison vous le faites, puisque votre but est essentiellement pratique.

Mais pour nous, une telle méthode aurait des conséquences gênantes.

Soit par exemple la forme *scriptorum*. Vous l'analyserez toujours en *scriptor-um* (génitif pluriel de *scriptor*), alors que dans bien des cas, ce sera le génitif pluriel du participe, pour lequel il faut couper *script-orum*. Il est possible que la désinence la plus brève offre l'analyse le plus fréquemment correcte, si bien que, pour des travaux d'intérêt pratique, cette simplification donne de bons résultats. Mais pour nous, elle risquerait de banaliser l'analyse et d'éliminer toute une série d'analyses correctes.

Busa

Dans le problème des degrés de probabilité de vérification des homographes, qui est un problème complexe, vous avez déjà choisi une ligne de travail pour votre recherche, n'est-ce pas? Mais le cadre d'une recherche complète est encore à définir: il y aurait encore plusieurs aspects différents à considérer, si on voulait poursuivre une recherche pure sur l'homographie de la langue latine, comme problème scientifique de langue.

Prosdocimi

Je me permets d'attirer l'attention sur le côté théorique de la polysémie - pour le latin on pourrait parler, sauf exceptions, d'homographie - qui a créé (au point de vue de la synchronie) de nombreuses difficultés concernant l'élaboration des théories linguistiques modernes. Au point de vue de la diachronie, on peut mieux expliquer la polysémie, et c'est d'après ce point de vue que je me permets d'attirer, comme je viens de le faire, l'attention sur l'éventualité de tirer profit de vos travaux préparatoires pour d'autres disciplines. Dans notre cas, une table, enregistrant tous les cas de polysémie d'un mot, peut faire prévoir certaines tendances évolutives, pourvu que l'on tienne compte que la langue fonctionne grâce aux distinctions, qu'elle est structurée et qu'elle évolue en acquérant sa structure, selon les principes économiques du moindre effort (dans le sens de A. Martinet).

C'est en me fondant aussi sur l'utilisation de votre analyse en différents domaines que je pensais à la possibilité d'étudier les rapports norme - système

(dans le sens de Hjelmslev et de Coseriu) et, en conséquence, la stylistique (dans le sens, par exemple, de l'italien Rosiello).

Je m'explique: dans le langage, nous avons d'un côté le système théorique, presque une borne de la formulation logistique et de l'abstraction, de l'autre l'usage qui représente la réalisation du système (naturellement le prius réel, c'est l'usage). Il faut ajouter à ces deux notions une troisième notion, la norme, que nous pourrions définir la projection sociale du système, c'est-à-dire la codification de certaines d'entre les nombreuses possibilités que le système offre, et d'autre que le système ne prévoit pas (encore). Je ne m'arrêterai pas ici sur l'importance de cette notion que je considère comme fondamentale pour la solution de certaines apories linguistiques, mais j'en soulignerai les implications stylistiques. Rosiello, s'inspirant de Coseriu, a défini le style du poète la possibilité de passer directement du système à l'usage, sans qu'il touche à la règle. Comme je crois que votre analyse, surtout au cours de la phase préparatoire, permet de vérifier les possibilités théoriques du système, nous avons donc lá un matériel de base très précieux pour des recherches de ce genre.

Une dernière curiosité: je n'ai pas beaucoup entendu parler des probabilités, dont on a reconnu les applications à la lignuistique très importantes (je cite, par exemple, les travaux de Herdan, de Guiraud, du sémitiste italien Fronzaroli): quels sont les points de contact éventuels entre votre travail et le leur?

Delatte

La plupart des textes latins sont de la prose d'art ou de la poésie. Ce n'est donc pas du tout le même problème que de traduire des articles de journaux. Dans un auteur comme Plaute, les amphibologies se lèvent avec facilité au fur et à mesure que la phrase progresse. La plupart du temps, le verbe n'est pas à la fin: il est dans la position logique, comme en français, et cela diminue les niveaux d'amphibologie. César, parce qu'il était essentiel qu'il soit compris, a lui aussi relativement peu d'amphibologies: les amphibologies se lèvent facilement au fur et à mesure de la lecture. Mais dans un auteur comme Tacite, la phrase devient d'une complexité telle, et l'auteur a accumulé lui-même tant d'amphibologies, que le lecteur est obligé de s'arrêter et de revenir en arrière.

Et c'est la raison pour laquelle, dans beaucoup de textes latins, on trouve, je crois, une telle difficulté de lever les amphibologies.

Stainier

Je voudrais poser deux questions au laboratoire de Liège: d'une part, pensez-vous pouvoir encore accélérer votre travail d'analyse philologique? Vous disiez pouvoir faire mille cartes par jour. Pensez-vous pouvoir encore accélérer cela en remplaçant une partie du travail postérieur du philologue par un travail de la machine? Je pose cette question en fonction de la considération suivante: votre but est directement philologique, littéraire; comme M. Delatte le disait, vous ne pouvez absolument rien laisser échapper comme phénomène philologique ou littéraire. Pour quelqu'un qui étudierait par exemple les textes philosophiques du moyen âge et serait moins directement intéressé par les problèmes d'esthétique littéraire, n'y aurait-il pas une accélération possible du travail comportant éventuellement certains risques quant à l'analyse esthétique?

Ma deuxième question porte sur les problèmes posés par le P. Busa, et qui se présentent peut-être plus rarement dans les textes classiques que dans les textes médiévaux; le problème, par exemple, où un mot, courant dans une signification, n'a que très exceptionnellement une seconde signification. Vous avez dit hier que, dans la "préédition" que vous faites de votre texte, vous sépariez les enclitiques. Ne croyez-vous pas qu'il y aurait intérêt à ce que ces mots à seconde signification très exceptionnelle soient, lors de la préédition, marqués d'un code particulier lorsqu'ils se rencontrent avec cette signification exceptionelle? Cela n'éviterait-il pas à la machine tout un travail de recherche, en même temps que cela attirerait immédiatement l'attention du philologue?

Evrard

En ce qui concerne la vitesse, la machine analyse environ huit cents mots par heure. Dès lors, ce qu'un philologue traite en un jour, est fourni par la machine en une heure ou un peu plus. Et c'est bien suffisant. On pourrait certes augmenter la vitesse en passant à une machine plus rapide, mais cela demanderait toute une transposition de programmes.

En revanche il me semble qu'on ne peut envisager qu'avec une extrême prudence un gain de vitesse obtenu par une réduction des lexiques ou par une limitation du nombre d'analyses présentées par l'ordinateur pour chaque mot. Sans doute, s'il s'agit de textes philosophiques du moyen âge, par exemple, il vous est loisible d'organiser le lexique que vous utilisez en fonction de cette langue. Je crois avoir expliqué, avant hier, qu'il peut nous arri-

ver de modifier le contenu des lexiques qui servent à l'analyse, par exemple en ajoutant des mots nécessaires pour tel ou tel auteur. Mais, de la même manière, nous pouvons supprimer des mots dont nous savons bien qu'ils ne sont pas employés par tel auteur.

Il en résultera une double accélération. Tout d'abord, plus le lexique est restreint, plus la vitesse de la machine augmente, parce que la démarche de consultation des lexiques est raccourcie d'autant. D'autre part, on évite ainsi de produire des cartes qui répondent à des analyses formellement possibles, mais dont on sait bien que, pour le genre de textes traités, elles sont non seulement très improbables, mais pratiquement impossibles.

Quant à la préédition, M. Delatte en a parlé à propos des enclitiques. Il faut ajouter que nous développons les nombres écrits en chiffres romains, et que nous leur donnons un code qui signifie que dans le texte, ce nombre est écrit en chiffres. C'est le traitement que nous appliquons aussi aux autres mots écrits en abrégé, par exemple les prénoms. Ce système peut être développé autant qu'on le veut et l'on peut fort bien prévoir un code qui signifierait, par exemple, mot technique de la logique, ou toute autre chose. Il faut toutefois ajouter que, dans la mesure du possible, le travail de préédition ne doit pas être trop alourdi.

En ce qui concerne le travail pratique de perforation des textes, il est fait au LASLA par une secrétaire qui ne connaît pas le latin, et cela ne pose pas de problème. Lorsqu'elle trouve un nombre écrit en chiffres romains ou un mot en abrégé, c'est elle qui nous demande comment il faut le développer. Comme, de toute manière, il y a toujours un contrôle philologique de la perforation, nous retrouvons les erreurs.

Le danger d'erreur le plus grand vient des lemmes rares qui ne sont pas encore dans le lexique. Si pour une certaine forme l'ordinateur ne trouve rien du tout, l'attention du philologue est attirée, puisque la ligne reste blanche: il y a une forme et une référence, mais il n'y a pas de lemme ni d'analyse. Le philologue sait donc qu'il doit compléter. Si la bonne analyse se trouve parmi celles que l'ordinateur a découvertes, il n'y a non plus aucun danger d'erreur. Mais si l'ordinateur présente une ou plusieurs analyses dont aucune n'est l'analyse voulue, le philologue peut être tenté d'adopter, mécaniquement, pour ainsi dire, l'une de celles qui sont présentées. C'est ainsi que, dans un index de César, la forme *indicta* a été par erreur rattachée à *indicere* parce que l'adjectif *indictus* (avec le préfixe négatif) n'était pas dans nos lexiques. Cet exemple vous montre pourquoi nous ne voulons pas d'un gain de temps, d'ailleurs minime, qui, obtenu par une réduction des lexiques, risque-

rait de conduire trop souvent à des erreurs. Nos études ne répondent pas à des besoins d'une urgence telle que nous puissions prendre ces risques d'erreur. La situation est toute différente pour les travaux d'intérêt pratique qui doivent être terminés dans un certain délai. Pour nous, tel n'est pas le cas. La seule raison d'être de nos travaux c'est de tendre à la perfection.

Busa

Il y a encore à dire quelques mots à propos de la polymorphie. Il y en a plusieurs types. Il y a la polymorphie due à des variations purement graphiques. Par exemple: *Heva* avec *h* initial, ou *Eva* sans *h* initial; *intelligo* et *intellego;* la diphtongue *ae* écrite comme *e;* ce qui fait quand-même une différence pour la machine, et aussi *conceptio, concepcio,* etc. Un autre degré de la polymorphie est à la limite entre la polymorphie et l'homographie: par ex., *animal, animalis* substantif, *animalis, animale* adjectif.

Nous avons établi des règles selon lesquelles nous avons réuni plusieurs lemmes sous un même numéro-code, en appellant ce phénomène "unification des lemmes". Par ex. le lemme Eva sans h, et le lemme Heva avec h, tous les deux ont le même code numérique. Par conséquent les formes seront mêlées dans un même paradigme, et non pas partagées sous des lemmes différents.

Je voudrais ajouter des cas curieux de lemmatisation impossible. De St. Thomas d'Aquin, comme vous savez, il y a l'autographe des trois livres du "Contra Gentes" et quelque peu du Commentaire sur "De Trinitate Boetii". De ces autographes, avec un travail énorme, les Pères Dominicains de la Leonine ont publié les mots que St. Thomas a effacés dans son manuscrit. Nous avons perforé et élaboré tous ces mots aussi. Parmi eux il y avait *modibus,* erreur dont il s'est aperçu à temps: il l'a effacée, et substituée par *modis.* Pour nous le problème a été: que faire de *modibus?* On l'a lemmatisé comme variation graphique de *modis.*

Un autre cas: un texte thomistique apocryphe dit: il y a des personnes qui ne prononcent pas *tempora, saltem,* mais *timpora, saltim,* qui n'est pas une forme correcte ni parlée, ni écrite.

Bodson

Est-ce que vous avez l'intention de développer la flexion totale du Forcellini dont vous venez de parler? Y êtes-vous décidé, ou considérez-vous que ce serait une chose intéressante à faire?

Busa

La faire n'est pas encore compris dans notre budget opérationnel immédiat: mais je pense qu'elle serait extrêmement intéressante, du moins sous l'aspect de science pure de la langue.