« Neophilologica ». Vol. 22 : *Études sémantico-syntaxiques des langues romanes*
Réd. Wiesław Banyś. Katowice 2010
Prace Naukowe Uniwersytetu Śląskiego nr 2828

Krzysztof Bogacki
Université de Varsovie

Les ressources lexicales dans les langues contrôlées

Abstract

This article deals with the lexicon, which is an important part of controlled languages. It is designed so as to increase readability of texts written using this subset of language by reducing their ambiguity and complexity. It is shown that developing a controlled language lexicon by severely reducing its size, thus eliminating some lexemes and approving only some of the possible meanings leads to conflicts. Very often these arise between general grammatical principles governing a controlled language and the terms to be included in the lexicon.

Keywords

Controlled languages, lexicon, polysemy, readability of texts, tools for controlled languages.

Introduction

Dans la structure des langues contrôlées (= LCs), un des éléments essentiels est le lexique. Cet article propose une réflexion sur cette composante dans la perspective de ses rapports avec la syntaxe, la morphologie et la stylistique en vigueur dans les LCs. Nous verrons que le contenu du lexique autorisé est partiellement fonction des exigences formulées à chacun de ces niveaux. Elles sont souvent contradictoires, agissent en sens inverse, et mettent parfois en échec les attentes des concepteurs des LCs. Nous aborderons aussi brièvement le problème du contrôle informatique de l'usage qui en est fait par l'utilisateur final.

Les langues contrôlées

Les langues contrôlées sont développées depuis peu de temps. Elles ont été créées dans les années ’30 du siècle dernier pour répondre à un besoin urgent qui se faisait sentir entre autres au sein de sociétés multinationales tenues d’assurer une parfaite connaissance de la documentation accompagnant produits et services dans le très vaste réseau international dans un contexte de personnes ne maîtrisant pas suffisamment la langue du message. Subsidiairement, elles ont été utilisées comme variantes simplifiées de langues standard et adoptées comme point de départ commode pour l’apprentissage des langues naturelles. Le cas le plus ancien, datant de 1932, est le BASIC English (acronyme de British American Scientific International Commercial English) créé par Charles Kay Ogden. Fondé sur la fréquence des mots en ce qui concerne le lexique et le temps moyen d’apprentissage des bases d’une langue, il a été utilisé pour la rédaction de manuels militaires pendant la seconde guerre mondiale. Dans le domaine français, on notera le « Français Fondamental » de G. Gougenheim élaboré dans les années ’50 et plus récemment le « Français Rationalisé » de Dassault Aerospace. Récemment, elles réapparaissent dans un tout autre contexte : comme outils pour la formulation d’alertes de toute sorte (attentats terroristes, émeutes, incendies, innondations ou autres catastrophes naturelles), pour la rédaction de tracts divers, modes d’emploi d’appareils, notices de médicaments etc., bref, dans de nombreuses situations où une compréhension rapide et correcte de messages est une priorité.

Quelles sont les caractéristiques les plus importantes des LCs ? La notion de LC implique d’un côté une composante proprement linguistique et de l’autre se rapproche d’une feuille de style imposant des conventions de bonne rédaction. Sur le plan strictement linguistique, il est plus facile de parler plutôt de tendances générales au niveau de la définition des standards des LCs que de donner une liste précise des traits qui les caractérisent (cf. S. O’Brien, 2003). Ces traits concernent la grammaire et le lexique.

Au niveau grammatical elles relèvent de la morphologie et de la syntaxe ; l’aspect phonétique, lui, n’est pas concerné étant donné que les LCs apparaissent principalement dans la version écrite. Les contraintes propres aux LCs s’expriment par l’élimination de certaines formes flexionnelles dont l’usage est strictement interdit (p.ex. de l’impératif qui doit être remplacé par l’infinitif) et par un certain nombre de recommandations positives (p.ex. remplacement des pronoms personnels par leurs référents nominaux) ou, au contraire, négatives (omission d’articles, élimination de la voix passive etc.). Les contraintes syntaxiques, quant à elles, concernent la longueur des phrases et des syntagmes, la complexité et le nombre de schémas syntaxiques admis. Elles sont souvent liées à des choix lexicaux qui conduisent à admettre un lexique « autorisé ». En effet, les LCs n’acceptent pas celui de la langue générale à cause de la polysémie qui y est omniprésente. Comme on sait,

elle est une des sources de l'ambiguïté qui nuit à la clarté des textes. Or celle-ci est une des exigences fondamentales qui se retrouve dans toutes les spécifications des LCs.

Les LCs ne sont pas faites pour rédiger des reportages, créer des essais littéraires ou encore moins d'écrire de la poésie. Certaines contraintes ont de très fortes répercussions négatives sur le plan stylistique. Ainsi la préférence donnée à la répétition de mots plutôt que l'emploi des synonymes ou encore le souci d'éviter les ellipses introduit une lourdeur et une monotonie considérables dans les textes en LC (contrairement à la consigne souvent invoquée : « Écrivez en bon français ! »).

Quelques malentendus relatifs aux LCs

Quelles que soient les caractéristiques des LCs, on s'aperçoit que parmi les traits recherchés se trouvent la clarté et la lisibilité qui découlent de la simplicité syntaxique (souvent associée à la brièveté du message) et la compréhensibilité ayant sa source dans l'univocité lexicale qui conduisent à une plus grande facilité de traduction.

> Controlled languages […] have been created in order to resolve problems of readability (reducing the complexity of syntactic structures of a text increases its readability), of comprehensibility (a lexical disambiguation increases the comprehensibility of a text) and of translatability (a syntactic and semantic control facilitates the shift between two languages) but not of grammaticality (a grammatical text written in a given CL will not necessarily be considered as grammatical in the corresponding natural language) (L. Spaggiari, F. Beaujard, E. Cannesson, 2003: 2).

Or une étude faite sur un corpus de textes concernant la protection anti-feu récoltés sur Internet et contrôlés selon les règles élaborées pour le polonais dans le cadre d'un projet européen[1] montre que les textes en LC ne sont pas toujours plus courts que ceux en langue standard. Bien au contraire, ils dépassent souvent les textes standards et la différence de longueur (calculée aussi bien en mots qu'en caractères) atteint parfois 25 %. L'explication est facile à donner : l'élimination des ellipses, le remplacement des pronoms par les groupes nominaux ou par les substantifs ajoute des mots et des caractères supplémentaires.

[1] Alert Messages and Protocols, MESSAGE, JLS/2007/CIPS/022.

Les langues contrôlées et la traduction[2]

D'autre part une recherche sur la traduction des textes contrôlés a montré que ceux-ci n'étaient pas plus faciles à traduire que les textes en langage standard. Ce malentendu vient probablement de l'observation que la réduction du lexique peut conduire à celle des schémas syntaxiques ce qui a priori crée de meilleures conditions pour la lisibilité, la compréhension et la traduisibilité des textes produits. Or, nous avons montré ailleurs (K. Bogacki, 2009) que les résultats de la traduction dépendent dans une grande mesure de la structure du dictionnaire de transfert : langue source / langue cible utilisé par le programme. La qualité du texte traduit est directement proportionnelle au nombre d'entrées avec un équivalent unique dans la langue cible dans le dictionnaire de transfert. L'idéal — irréalisable, cf. infra — consisterait à disposer d'une liste de lexèmes bi-univoque. Cette observation montre l'importance que revêt le lexique autorisé dans la perspective multilingue et unilingue.

Le lexique autorisé

Les restrictions les plus importantes imposées sur la composante lexicale des LCs sont d'ordre qualitatif : exclusion d'entrées polysémiques, d'homonymes, de synonymes, d'archaïsmes, de régionalismes aussi bien dans une perspective unilingue que multilingue (dictionnaire de transfert). Ces restrictions entraînent une forte diminution du nombre d'entrées ce qui est d'ailleurs conforme à l'idée générale des LCs. Elle est très sensible lors de la constitution d'un lexique unilingue mais reste relativement moins visible dans un dictionnaire de transfert. En effet, l'élimination de sens dans les entrées du lexique source et dans le lexique cible ne se fait pas dans les mêmes proportions. Il arrive souvent que différentes acceptions soient rendues par un même équivalent. Ainsi *elektryczny* attesté en polonais avec trois sens (**1.** «odnoszący się do elektryczności — energii elektrycznej» 'se rapportant à l'électricité', **2.** «zasilany elektrycznością» 'alimenté par l'énergie électrique', **3.** «przewodzący prąd elektryczny») est rendu chaque fois en français par *électrique*.

La réduction du nombre d'entrées lexicales est confirmée lorsqu'on examine différentes LCs existantes. Ainsi dans la version de l'anglais contrôlé proposée par McDonnell Douglas Corp. en 1979, on trouve 2000 mots, le BASIC English d'Ogden n'en contient que 850 (400 noms d'ordre général, 200 illustrés, 150 ad-

[2] Cf. W.O. Huijsen (1998a, 1998b) ; K.M. Stewart (1998).

jectifs qualificatifs dont 50 opposés + 100 termes internationaux et 50 termes spécifiques)[3]. Parfois on divise le lexique autorisé en 2 sections : celle de base, propre à toutes les variétés de LC (protection civile, toxicologie, aéronautique etc.) et celle qui est spécifique au domaine traité. En ce qui concerne l'étendue du lexique de base, elle est restreinte. Grâce à la jonction avec une base de données lexicales le nombre total d'items autorisés peut être très élevé. Cette solution est d'autant plus justifiée qu'il serait impossible de vouloir remplacer les termes techniques relevant de la terminologie par des structures lexicales nouvelles formées d'items contenus dans le lexique de base. D'autre part ces termes nouveaux, forgés pour les besoins du moment, auraient le statut des composés et devraient figurer comme tels comme entrées distinctes dans un dictionnaire à part.

Voyons quelques problèmes qui surgissent au moment de la construction du lexique de la LC limité à la protection et à la lutte anti-feu[4].

Dans la perspective unilingue, le plus souvent le choix d'un lexème se joue sur le plan sémantique. En premier lieu, on élimine tout ce qui conduit à la polysémie. Dans un lexème ayant plusieurs sens on n'en autorise qu'un, celui qui est propre au domaine traité. Or bien rares sont les entrées lexicales monosémiques qui ne nécessitent aucune intervention et peuvent être transférées du lexique général sans aucune restriction. Dans une liste de 561 lexèmes trouvés ils sont à peine 27 c'est-à-dire forment moins de 5 %. Ce sont pour la plupart des termes techniques. Ainsi dans le SJP on trouve :

Acetylen 'gaz palny, stosowany m.in. do spawania i cięcia metali' ; 'carbure d'hydrogène insaturé, à triple liaison, gaz combustible, utilisé pour la soudure des métaux' ;

azbestowy (adjectif tiré de *azbest* 'minerał o budowie włóknistej, odporny na działanie wysokiej temperatury i na ścieranie' ; 'espèce de minéraux (groupe des *Silicates*) à structure filamenteuse, assez souple et résistante, relativement incombustibles') ;

gaśniczy 'stosowany do gaszenia pożaru' ; 'utilisé pour combattre l'incendie'.

Le cas de loin le plus fréquent est cependant celui des entrées polysémiques. Tel est le cas du verbe *być,* présenté comme suit par le SJP :

być (verbe plein)
1. «istnieć, żyć» 'exister'
2. «być obecnym, znajdować się gdzieś» 'se trouver quelque part'

[3] Cf. B. Gingras (1987).

[4] Les recherches sont basées sur 10 textes trouvés sur Internet ou affichés dans des chambres d'hôtels. Ils ont été rédigés principalement par les services de pompiers à l'intention du grand public.

3. «trwać przez pewien czas, zdarzać się, odbywać się» ‘se produire, avoir lieu’

4. «brać w czymś udział» ‘participer, prendre part’, comme dans *Ojciec był w powstaniu* ‘Le père a pris part à l’insurrection’

5. «uczęszczać gdzieś, korzystać z czegoś » ‘aller à’, p.ex. *Dwa lata był w technikum.* ‘Il était pendant deux ans dans un lycée technique’

6. «mieć jakieś wydarzenia za sobą, spodziewać się ich lub być w trakcie jakichś wydarzeń » ‘être en cours de’, p.ex. *Gdy przyszli goście, byliśmy już po obiedzie.* ‘Quand les invités sont venus, nous avons déjà mangé’

7. «sięgać, dostawać do jakiejś wysokości » ‘aller jusqu’à’ p.ex. *Woda była do kostek* ‘L’eau allait jusqu’aux chevilles’

8. «znajdować się w jakimś stanie, ulegać czemuś» ‘subir l’influence de, être dans un état de’, p.ex. *Był pod wpływem alkoholu.* ‘Il était sous l’emprise de l’alcool’

auxquels il conviendra d’ajouter 3 sens « auxiliaires » supplémentaires :

być (auxiliaire)

1. « verbe servant à former les formes composées de certains temps (futur analytique) et utilisable à la voix passive »

2. « verbe copule », p.ex. *Był wysokiego wzrostu.* ‘Il avait une grande taille’

3. « dans des tournures impersonnelles », p.ex. *Trzeba było to zrobić.* ‘Il fallait le faire’

Les textes que nous avons examinés invitent à éliminer 7 sens non-attestés : le sens existenciel (1), celui de participation (4, 5), et les sens (3, 6, 7, 8) pour ne retenir que le sens locatif (cf. *Gaśnice są na każdym piętrze* ‘Les extincteurs sont à chaque étage’). Parmi les emplois « auxiliaires » seule la valeur de copule est souvent attestée comme dans *Jeśli ranny jest nieprzytomny* ‘Si le blessé est sans connaissance’. Il a donc au départ deux valeurs à retenir pour un même verbe ce qui pose le problème de la monosémie soulevé avec insistance par les théoriciens des LCs. Si l’élimination de la valeur « copule » semble impossible, on sera obligé de postuler pour l’expression de la localisation, exprimée tout naturellement par *być,* de recourir à *znajdować się,* plus neutre que *leżeć* ou à un autre verbe synonyme. L’inclusion d’un de ces mots dans la liste des entrées autorisés empêchera automatiquement de s’en servir avec un autre sens qu’il est susceptible d’exprimer.

Il arrive souvent que l’on opère des choix qui ne sont pas nécessairement ceux des lexicographes qui au moment de rédiger une entrée de dictionnaire tiennent compte de la fréquence d’emploi. Considérons l’adjectif *agresywny* décrit par le SJP comme suit :

1. «zachowujący się wrogo, napastliwie» ‘qui a un comportement hostile’

2. «pełen agresji» ‘plein d’agressivité’
3. «stosujący przemoc» ‘porté à la violence’
4. «pełen ekspresji, dynamiki» ‘plein d’expressivité et de dynamisme’
5. «o kolorze: jaskrawy, ostry» ‘offensant le bon goût par son caractère excessif et provocant’
6. «bardzo aktywny chemicznie» ‘chimiquement très actif’

Parmi les 6 sens associés à cet adjectif, le seul attesté dans le domaine de la protection anti-feu figure en dernière position dans la liste du SJP (‘chimiquement très actif’). Il est probablement le moins fréquent dans les textes globalement.

La création du lexique autorisé dans une perspective multilingue ne fait que compliquer les choses. Il s’agit, dans ce cas-là, de tenir compte dans la plus grande mesure possible des deux langues à la fois et de respecter autant qu’il se peut les règles grammaticales et les principes généraux de LCs.

Or, il arrive que la réalité langagière impose de violer le principe de la correspondance bi-univoque entre langue source et langue cible. Ainsi le substantif polonais *rzeka* devrait être associé à deux substantifs français : *fleuve* et *rivière* pour éviter le risque de mauvaise traduction (p.ex. *rzeka* par *fleuve* au lieu de *rivière* ou inversement, selon le cas). La situation n’est pas exceptionnelle, cf. *stryj/wuj* ‘oncle’, *wujenka/stryjenka* ‘tante’.

Il arrive que des facteurs autres que sémantiques influencent le choix du lexème pour le lexique autorisé[5]. Or le souci d’éviter l’ambiguïté qui doit être placé très haut dans la hiérarchie des consignes entraîne des conséquences sur le plan du lexique autorisé. Considérons la phrase

Jeśli wyciek acetylenu poprzedza odcięcie zasilania…

Elle est ambiguë entre

‘Jeśli wyciek acetylenu występuje przed odcięciem zasilania’ = ‘Si la fuite d’acetylène précède la coupure de l’alimentation’

et

‘Jeśli odcięcie zasilania występuje przed wyciekiem acetylenu’ = ‘Si la coupure de l’alimentation précède la fuite d’acetylène’

Pour lever l’ambiguïté, on proscrira *poprzedzać* ‘précéder’ au profit de *następować po* ‘suivre’ :

[5] Cette situation, très fréquente dans la langue, est liée à l’homonymie des formes grammaticales : entre les formes nominales (substantif / substantif, substantif / adjectif, adjectif / adjectif), à l’intérieur du système verbal entre différentes classes de verbes, et entre formes nominales et verbales. Cf. E. Awramiuk (1999) qui en a étudié plus de 130 types.

→ *Jeśli wyciek acetylenu następuje po odcięciu zasilania*
→ *Jeśli odcięcie zasilania następuje po wycieku acetylenu*

Une autre solution consisterait à autoriser le passif banni pourtant pour des raisons incompréhensibles par la majorité de spécifications des langues contrôlées :

Jeśli wyciek acetylenu jest poprzedzony przez odcięcie zasilania

Le lexique des LCs contrôlé par des outils informatiques

Les langues contrôlées, destinées à un public humain, sont dépourvues d'éléments formels qui faciliteraient le recours à des outils informatiques. Ceux-ci seraient pourtant bienvenus. En effet, la rédaction de textes contrôlés semble un exercice fastidieux mais peut être facilement assistée par des outils informatiques relativement simples. Or, il existe des éditeurs intégrant plus ou moins de règles de la LC. Pour le français, on dispose du Compagnon LiSE[6] qui, à chaque entorse au code de bonne formation du texte, affiche les règles explicitement formulées dans le guide de rédaction. Pour les resources polonaises, l'outil du rédacteur CONTROLEDIT élaboré dans le cadre du projet MESSAGE est plus sobre. Il est pourvu d'un module lexical contenant les mots autorisés consulté par le programme lors de la saisie du texte[7]. Toute tentative d'insertion d'une forme non-reconnue par le module (aussi bien au point de vue lexical que grammatical) provoque un avertissement qui invite l'usager à reformuler le texte. Le module n'a pas de pouvoir contraignant en ce sens que l'emploi de forme non-autorisée ne provoque pas le blocage du programme qui empêche l'utilisateur de passer à l'étape suivante. CONTROLEDIT utilise simultanément deux ressources lexicales : le dictionnaire morphologique général, crypté et compilé, faisant partie inamovible du programme lui-même et le dictionnaire de spécialité propre au domaine traité. Le mérite de CONTROLEDIT consiste en la possibilité d'ajouter les modules linguistiques très différents préparés par l'utilisateur : circulation routière, services d'urgence des centres hospitaliers, émeutes possibles durant les manifestations sportives etc.

[6] Voir le projet ANR LiSE : *Linguistique, normes, traitement automatique des langues et Sécurité : du Data et Sense Mining aux langues contrôlées*. http://projet-lise.univ-fcomte.fr/.

[7] Il est clair cependant que le look-up ne concerne que le côté formel et non pas sémantique des mots saisis.

Conclusions

Prises une à une, les règles qui régissent le lexique des langues contrôlées vont dans le sens de la simplification du système total de ce sous-ensemble de la langue et devraient correspondre parfaitement aux exigences imposées aux LCs par leurs concepteurs et théoriciens. En réalité cependant elles provoquent des conflits avec d'autres exigences ou se concurrencent et s'excluent mutuellement. Le résultat final apparaît moins encourageant que prévu.

Références

Allen J., 1999: "Different Types of Controlled Languages". In: *Technical Communicators Forum TL 1*. http://www.tc-forum.org/topiccl/cl15diff.htm.

Awramiuk E., 1999: *Systemowość polskiej homonimii międzyparadygmatycznej*. Białystok, Wydawnictwo Uniwersytetu w Białymstoku.

Bogacki K., 2009: "Controlled Languages and Machine Translation". In: *International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains*. Presses Universitaires de Franche-Comté, 69—73.

Bogacki K., 2010a : « Les langues contrôlées et les traits sémantiques ». Dans : T. Giermak-Zielińska, A. Dutka-Mańkowska, éds. : *Des mots et du texte aux conceptions de la description linguistique*. Warszawa, Wyd. UW, 38—44.

Bogacki K., 2010b : « Sur quelques malentendus relatifs aux langues contrôlées ». Dans : J. Górnikiewicz, H. Grzmil-Tylutki, I. Piechnik, éds. : *En quête de sens. W poszukiwaniu znaczeń. Études dédiées à Marcela Świątkowska. Studia dedykowane Marceli Świątkowskiej*. Kraków, Wyd. UJ, 103—111.

Gingras B., 1987: "Simplified English in Maintenance Manuals". *Journal of the Society for Technical Communication*, **34**(1), 24—28.

Gougenheim G., Michéa R., Rivenc P., Sauvageot A., 1956 : *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris, Didier.

Huijsen W.O., 1998a: *Completeness of compositional translation*. Utrecht : Drukkerij Elinkwijk B.V.

Huijsen W.O., 1998b: "Controlled language : an introduction". *CLAW*, **98**, 1—15.

O'Brien S., 2003: "Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets". *EAMT-CLAW* www.mtarchiveinfo/CLT-2003-Obrien.pdf.

Ogden C.K., 1930: *The Basic Words*. London.

Spaggiari L., Beaujard F., Cannesson E., 2003: "A Controlled Language at Airbus". *EAMT-CLAW* http://www.mt-archive.info/CLT-2003-Spaggiari.pdf.

Stewart K.M., 1998: *Effect of AECMA. Simplified English on the comprehension of aircraft maintenance procedures by nonnative English speakers*. PhD, University of British Columbia.

http://projet-lise.univ-fcomte.fr/
http://sjp.pwn.pl/ (=SJP)
http://spellchecker.ru/English-German-Translation/English-French-Translation