Journal des traducteurs
Translators' Journal

Dictionnaires automatiques pour traducteurs humains

Lydia Hirschberg

Volume 10, numéro 3, 3e trimestre 1965

Traduction automatique et informatique

URI : https://id.erudit.org/iderudit/1061157ar
DOI : https://doi.org/10.7202/1061157ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0316-3024 (imprimé)
2562-2994 (numérique)

Découvrir la revue

Citer cet article

Hirschberg, L. (1965). Dictionnaires automatiques pour traducteurs humains. *Journal des traducteurs / Translators' Journal*, *10*(3), 78–86. https://doi.org/10.7202/1061157ar

DICTIONNAIRES AUTOMATIQUES POUR TRADUCTEURS HUMAINS*

Lydia HIRSCHBERG,
Centre de linguistique automatique appliquée,
Université libre de Bruxelles

L'analyse du néerlandais écrit, effectuée en vue de cette application, se limite à une réduction morphologique et à une reconnaissance de structures définies à une transformation près. Mais elle nous fournit quelques données sur des notions d'unité lexicale et de grammaire transformationnelle, spécifiques pour ces problèmes particuliers.

Le cadre pragmatique dans lequel nous nous trouvons est celui de la construction de dictionnaires automatiques de locutions pour traducteurs humains, français-néerlandais et néerlandais-français, réalisés au Centre de Linguistique Automatique Appliquée de l'Université de Bruxelles en collaboration avec le Bureau de Terminologie de la CECA à Luxembourg, dirigé par Monsieur J. A. Bachrach [1].

Ces travaux font suite à ceux qui ont mené à la réalisation du « Dicautom », dictionnaire automatique pour traducteurs humains [2].

Nous avons appelé ces nouveaux dictionnaires, des *dictionnaires de locutions,* parce que le traducteur peut leur demander une série de phrases écrites pour lesquelles il éprouve des difficultés.

◆ Le programme doit comparer une phrase demandée avec des phrases-exemples consignées dans son dictionnaire et trouver la ou les phrases qu'il possède et qui sont les « plus proches » possibles du texte demandé. A défaut de phrases trouvées, il répond en donnant des traductions d'unités lexicales isolées, de sorte que le contenu de ce dictionnaire automatique est très semblable au fichier personnel d'un traducteur.

*
* *

---

* Exposé fait le 18 février 1965 aux *Journées de Linguistique Appliquée,* organisées par l'Association française de linguistique appliquée. Ce travail a été réalisé grâce à un contrat avec la Haute Autorité de la *Communauté Européenne du Charbon et de l'Acier* (Luxembourg). Nous remercions Mme Hirschberg pour la permission qu'elle nous a accordée de reproduire cet article.

1 — Bachrach, J., Decresy, Fr., Goetschalkx, L., Hirschberg, L. et Van Beek, H., *LOCFRA : Dictionnaire des locutions françaises* (Nouvelle version du DICAUTOM), sous presse, 1965.

Decresy, Fr., Hirschberg, L. et Van Beek, H., *Analyse morphologique du néerlandais.* Université Libre de Bruxelles (sous presse), 1965.

2 — Bachrach, J. A., Blois, J., Decresy, Fr., Hirschberg, L. et Mommens, J., "DICAUTOM", *La Traduction automatique,* IV.3 (1963).

Nous avons parlé des phrases « les plus proches » d'un texte demandé. Cette proximité est définie de la manière suivante, à l'aide d'un calcul de valeurs de coïncidence, sur lequel nous reviendrons.

¶ Deux fragments de texte sont réputés d'autant plus *proches* qu'un plus grand nombre de leurs *unités lexicales coïncident,* et cela, *sans tenir compte*

a) de *l'ordre* dans lequel elles se présentent;

b) du fait qu'elles ont *la même forme* ou *deux formes différentes,* parmi celles que l'unité en question est habilitée à prendre tout en conservant les mêmes lois de traduction, c'est-à-dire parmi celles qui appartiennent à une même rubrique;

c) du fait que les unités du groupe *se suivent immédiatement* ou sont *séparées* par d'autres;

d) des *mots « noirs »* qui pourraient s'intercaler dans la séquence, les mots noirs étant *ceux auxquels les constructeurs n'ont pas donné d'entrée dans le dictionnaire.*

Nous avons à montrer en quoi une telle notion de proximité est utile, mais il est bon de faire remarquer qu'elle admet la transformation la plus générale pour un groupe d'unités lexicales. En effet, elle admet deux groupes comme identiques à n'importe quelle transformation près, qu'elle soit paradigmatique ou positionnelle. Et nous arrivons ainsi à une limite de ce que l'on pourrait appeler une grammaire transformationnelle, limite utile pour notre application.

Ainsi dans notre système, la phrase: « *Nos hommes formaient des unités de combat d'élite* » est interchangeable avec les phrases suivantes:

- *Nos hommes d'élite furent formés en unités de combat.*
  ou
- *Nous avons proposé nos hommes pour former des unités de combat d'élite.*
  ou
- *Des unités de combat furent formées avec des hommes d'élite.*
  ou encore
- *A l'exception d'une minorité qui formait une unité de combat d'élite, la plupart de nos hommes se sont enfuis devant la soudaine contre-attaque,* etc...

N'importe laquelle de ces phrases, si elle était introduite dans le dictionnaire, comme phrase-exemple, permettrait de retrouver chacune des autres avec une coïncidence valant 52, somme des valeurs des coïncidences des quatre unités lexicales communes:

homme (valeur de coïncidence 1)
former (" " " 1)
unité de combat (" " " $30 = 3 \times 10$)
d'élite (" " " $20 = 2 \times 10$).

On remarquera que nous attribuons une valeur de coïncidence 10 à chaque fragment contenu dans une unité lexicale composée de plusieurs fragments séparés par des espaces.

Ces valeurs ont été choisies pour des raisons de commodité d'écriture et d'efficacité du programme qui fonctionne pour une machine (IBM 1410). Aucune considération fondamentale n'a présidé à ce choix, qui donne un moyen, parmi d'autres, pour privilégier l'unité lexicale la plus longue, lorsque certains de ses fragments forment eux-mêmes des unités indépendantes existant également dans le dictionnaire.

Mais il est clair qu'une aussi grande liberté doit être tempérée par une définition prudente de la notion d'*unité lexicale*, de manière à éviter certains accidents et qu'une définition purement distributionnelle comme celle du *morphème* de Harris [3] — qui n'est d'ailleurs pas un moyen pratique pour découvrir une unité intéressante pour quoi que ce soit — ne fournirait pas un critère suffisant.

On trouve, bien entendu, des cas très clairs où une caractérisation distributionnelle et sémantique permet de délimiter avec certitude un morphème ou unité lexicale sur une langue envisagée isolément. Pottier [4] nous en a donné un exemple en identifiant comme morphème en français « de ce côté » parce qu'il est synonyme du morphème « là » et qu'il est interchangeable, dans certains contextes, avec ce morphème; par exemple, dans la phrase:

« je vais | de ce côté ».
« je vais | là ».

En effet, à d'autres morphèmes, on ne peut juxtaposer « je vais de », mais seulement « je vais à ». L'expérience montre cependant que des cas aussi clairs sont exceptionnels.

A notre avis, ce n'est pas non plus dans la seule amélioration de la description physique de *l'unité de sens* que l'on trouve une solution: qu'elle puisse être « disjointe », comme le monème de Martinet, cela est bien évident [5]. L'important est de noter que ces linguistes, qui travaillent sur une langue isolée, aboutissent soit à une définition purement structurale, mais qui décrit une classe d'unités sans intérêt, soit une définition basée partiellement sur des considérations sémantiques, imprécises parce que foncièrement subjectives.

Notre unité lexicale est une notion non pas absolue, liée à une analyse du sens, mais une unité relative à une langue d'arrivée, remplaçant la notion imprécise de sens par la *notion opérationnelle de traduction*, indispensable pour un traitement automatique.

Par exemple, on parle dans le vocabulaire de l'urbanisme de « *Maisons construites en bandes* » ou simplement de « *Maisons en bandes* », ce qui signifie qu'elles sont construites *en une rangée*. De sorte que nous prendrons pour nouvelle unité lexicale */ en bandes /*, adjectif invariable d'ail-

3 — Harris, Zellig, "From Phonem to Morphem", *Language* 31 (1935).

4 — Pottier, Bernard, "Applications pédagogiques de la linguistique moderne", Journées de Linguistique appliquée, Université libre de Bruxelles, 27 février 1965.

5 — Martinet, André, *Eléments de linguistique générale*. Paris, Armand Colin, 1960.

leurs, parce qu'elle se traduit par / *in rijen* / et doit donc avoir une entrée, une rubrique de dictionnaire, différente du substantif féminin *bande,* traduit par « band, etc. ». Nous évitons ainsi que la présence et la coïncidence fortuite d'autres unités ne conduise à la sélection de phrases-exemples du dictionnaire, contenant l'une, alors que l'expression demandée contenait l'autre, indispensable pour informer correctement sur la traduction à fournir.

En bref, nous avons ici l'occasion d'une *définition pragmatique de la notion d'unité lexicale.* Cette notion est particulièrement étalée en néerlandais, relativement au français, autour d'un optimum qui serait une suite de signes entre deux espaces et que nous appelons *mot* pour la facilité de l'exposé. Ce mot est une unité sans intérêt linguistique, mais d'une grande importance dans les opérations automatiques parce qu'il est facile à déceler et à isoler. Il est particulièrement privilégié dans la langue écrite, la seule qui nous préoccupe ici. Or notre conception de la langue écrite doit être totalement indépendante de la langue parlée, malgré que les linguistes n'attachent de caractère fondamental qu'à cette dernière.

*
* *

Le néerlandais, d'une part, admet une composition qui ne se limite pas seulement à des préfixes ou des suffixes, appartenant à des listes courtes « fermées ». Quoique dans une mesure moindre que l'allemand, il existe en néerlandais une certaine composition libre de mots, appartenant à des listes longues, de sorte qu'il arrive qu'un mot ne possède pas d'entrée dans le dictionnaire, mais se scinde en plusieurs unités lexicales.

Nous éprouvons ici immédiatement le besoin de définir les limites [6] de ce qu'est la composition libre, d'un point de vue absolument différent de celui auquel se placent les traités classiques que nous avons consultés [7].

¶ Nous dirons que des *structures* sont *composées librement en néerlandais* (ou en allemand) *relativement au français,*

- *si* chaque composant conserve sa traduction comme s'il était seul,
- *et si* la structure de la traduction du composé est prévisible et peut être déduite de la nature des composants, mais ne dépend pas des composants particuliers que l'on traduit.

Par exemple, *Rue d'Ulm* devient *Ulmstraat,* composition entièrement libre.

---

6 — Kukenheim, L., "Van Glossarium tot Thesaurus", *Levende Talen* 203 (1959).

7 — Bally, Ch., *Linguistique générale et linguistique française* Berne, Francke, 2e éd., 1944, p. 352-363.
de Wooys, C. G. N., *Nederlandse Spraakkunst.* Groningen, Wolters, 1963. 6e druk (Woordvorming, p. 177-261).
van Haeringen, C. B., *Neerlandica.* Den Haag, Deamen N. V., 1962.

Nous avons de même :

- *expansie* = « expansion » ; *politiek* = « politique » ;
  expansiepolitiek = « politique d'expansion ».
- *zak* = « poche » ; *geld* = « argent » ; *zakgeld* = « argent de poche ».
- *fabriek* = « usine » ; *directeur* = « directeur » ;
  fabrieksdirecteur = « directeur d'usine ».
- *Woordenrijkdom* { *woord (en)* / *rijkdom* } → « richesse de mots »

Ce dernier exemple est déjà à la limite, et il vaudrait mieux introduire le composé comme unité lexicale dans le dictionnaire de manière à lui attribuer une traduction plus conforme à l'usage (« richesse de vocabulaire »), qu'il est impossible de déduire immédiatement des composants.

Il en est de même pour :

- *sport* = « sport » ; *nieuws* = « nouvelles » ;
  sportnieuws = « nouvelles sportives », et non pas « du sport ».
- *wereld* = « monde » ; *politiek* = « politique » ;
  wereldpolitiek « politique mondiale », et non pas « du monde ».

Enfin, voici des exemples qui sembleraient formés de trois mots :

- *steen* = « pierre » ; *kool* = « charbon » ; *mijn* = « mine » ;
  steenkoolmijn = « mine de houille », et non pas « de charbon de pierre ».
- *brand* = « incendie » ou « combustion » ; *stof* = « matériau » ;
  besparing = « économie » ;
  brandstofbesparing = « économie de combustible ».

Il se présente donc des cas où l'on peut faire la coupure, des éléments dont la recomposition est évidente au moment de la traduction. Mais il en est d'autres où pareille coupure est impossible ; c'est le cas pour *brandstof* et *steenkool*, qui doivent figurer en entier, comme unités lexicales indépendantes dans le dictionnaire, malgré leur formation régulière en néerlandais [8].

En effet, nous ne nous préoccupons ni d'étymologie, ni d'évolution de la pensée ou de l'usage par associations. La régularité même d'une structure dans un « composé » néerlandais n'est pas pour nous une condition suffisante pour l'exclure comme unité lexicale indécomposable. On trouve en effet très fréquemment de ces cas où la décomposition des éléments néerlandais paraît évidente, mais où les traductions obtenues ainsi sont inadmissibles et inutilisables.

Ainsi *staathuishoudkunde* = « économie politique » donnerait par décomposition « art de la tenue de la maison de l'état », tandis que *huishoudkunde* signifie « économie domestique » et est également impropre à la dé-

8 — On remarquera que ces exemples néerlandais se laissent facilement transposer en anglais, ce qui rend la démonstration valable pour cette dernière langue. Nous pensons même qu'elle est d'autant plus intéressante qu'elle part d'un domaine peu familier à nos lecteurs et prend ainsi une valeur pédagogique. *NDLR.*

composition à cause du résultat que l'on obtiendrait en français et qui n'aurait aucune valeur pratique.

D'autre part, certaines unités lexicales sont des groupes composés de plusieurs mots séparés par des espaces, juxtaposés, compacts, ou parfois séparés par d'autres unités, qu'il est nécessaire d'indiquer comme telles dans le lexique, parce que la traduction de l'ensemble n'est pas une transformation générale ni une simple juxtaposition des traductions de chacunes des parties, lorsque ces parties sont seules ou dans d'autres contextes.

Tel est le cas du groupe */aan de lopende band/* qui se dit de la *production* et se traduit par « *à la chaîne* » (littéralement « à la bande courante »). Tandis que l'acier « recuit » devient un acier qui a été rendu incandescent, */gegloeid staal/*.

Les multiples manières dont nous devons découper une langue en morphèmes, en fonction de chacune des langues vers lesquelles on traduit, tient en fait à la diversité même des phénomènes qu'il faut décrire et aux différences d'éclairage de langue à langue. Vinay distingue ainsi une classe de *modulations sémantiques figées,* qu'il illustre par l'exemple du *col de cygne* (sorte de microphone utilisé dans les laboratoires de langue) devenu « nuque d'oie » en anglais (*goose neck*) [9].

Il nous faut encore une fois nous débarrasser de toute considération sur l'histoire de la formation d'une unité, pour ne voir que son usage, en l'occurrence sa traduction, de sorte que rapprocher */chaîne/* et */à la chaîne/* devient pour nous *aussi absurde* que de rapprocher */sec/* et */section/* ou */but/* et */début/*.

La question se pose alors de savoir s'il faut n'introduire dans le dictionnaire comme unités lexicales que des groupes de mots entre lesquels rien ne peut s'intercaler et pour lesquels on n'admet qu'une certaine variation morphologique des termes mêmes, ou s'il faut admettre des transformations plus générales, sans toutefois arriver à la liberté que nous avons admise pour les phrases servant d'exemples.

Nous nous trouvons en fait devant l'obligation de résoudre dans la pratique le problème soulevé par Pottier [10] pour des groupes tels que *machine à coudre, crise de croissance,* etc., qui deviennent de moins en moins liés.

¶ Au départ du français, nous nous limitons empiriquement, pour l'introduction de têtes de rubriques dans le dictionnaire, aux *unités compactes.*

Ainsi nous n'introduirions pas */crise de croissance/* puisqu'on peut trouver *crise très... très aiguë de croissance.*

De même nous introduisons */à la chaîne/* mais non pas *production à la chaîne* parce que */production/* et */à la chaîne/* forment deux unités qui

---

9 — Vinay, J.-P., "Langage et découpage de la réalité", Journées de Linguistique appliquée, Université libre de Bruxelles, 27 février 1965.

10 — Pottier, B., "Vers une sémantique moderne", *Travaux de Linguistique et de Littérature,* Strasbourg II (1964).

Pottier, B., "Introduction à l'étude des structures grammaticales fondamentales", *La Traduction Automatique* III.3 (1962).

peuvent être séparées par d'autres et même être inversées dans l'ordre linéaire de la phrase. On peut avoir en effet: « *A la chaîne,* une telle *production* s'avère, etc . . . ». Dans ce cas, nous comptons sur la coïncidence de ces deux unités séparées pour que le programme découvre l'exemple le plus adéquat pour répondre à une demande, mais qui supporte la transformation la plus générale décrite plus haut.

Incidemment, cette liberté de transformation que nous avons soulignée est indispensable pour limiter le nombre de phrases-exemples introduites dans le dictionnaire, qui sans cela deviendrait impraticable; et aussi pour donner très librement n'importe quelle forme à une phrase qui sert d'exemple et qui doit permettre de retrouver n'importe quelle demande pour laquelle cet exemple est le plus adéquat, à l'aide d'un programme simple et rapide. Ainsi */Maisons en bandes/* doit pouvoir retrouver « *Maisons situées sur des collines exposées au sud, en bandes* . . . » et réciproquement.

¶ A partir du néerlandais cependant, cette limitation n'est pas admissible. Nous devons prévoir comme unités lexicales indépendantes des groupes entre lesquels plusieurs autres unités doivent prendre place librement.

Il suffit de rappeler qu'un verbe dit séparable se compose parfois d'un préfixe constitué par un substantif ou une « particule », collée à la forme infinitive ou au participe, mais qui s'en détache pour les formes conjuguées et va alors se placer quelque part dans la phrase, obéissant à des règles de position relativement libres, plus libres qu'en allemand où cette partie séparable ne va cependant pas se placer au bout de la phrase aussi souvent que le voudraient les grammaires.

Prenant un texte de Malraux dédicacé à du Perron et admirablement traduit par cet essayiste hollandais, je trouve par exemple:

« ses traits . . . n'*exprimaient* de la fatigue »

« zijn trekken . . . *drukten* niets (*uit*) dan moeheid ( ) »

La particule (*uit*) aurait pu voyager à travers ce texte et se mettre par exemple plus loin, à l'endroit indiqué par la parenthèse vide.

Ailleurs, la particule ne peut reculer, en particulier au-delà d'un autre infinitif, comme par exemple dans

« *Il s'efforçait de ralentir sa marche* »

« Hij *spande zich (in)* langzamer te lopen ».

Le critère idéal que nous proposons d'employer en néerlandais pour savoir si nous devons introduire un groupe locutionnel comme unité lexicale ou non serait alors celui-ci:

¶ *Si les phrases-exemples qui contiennent le groupe entier sont à exclure en tant qu'exemples pour* traduire des phrases qui *ne contiennent que des mots isolés* du groupe ou réciproquement, nous ferions d'un tel groupe une *nouvelle unité lexicale.*

Ce critère est classique en documentation automatique.

Dans la pratique, cependant, ces groupes locutionnels acceptent des transformations beaucoup plus réduites que la transformation très générale

que nous avons admise entre des unités indépendantes pour laquelle une phrase ainsi transformée reste encore un exemple valable de la phrase non transformée.

Or vouloir déceler, par un programme d'analyse plus complet, si les limites particulières permises pour la transformation de toutes les unités lexicales non compactes ont été atteintes ou non, n'est pas une solution réaliste, parce que cela ralentirait considérablement la recherche sur la petite machine sur laquelle nous travaillons, sans utilité essentielle pour notre problème.

Nous nous limitons donc aux locutions non compactes pour lesquelles les règles d'identification s'appliquent aisément et sont indispensables, comme c'est le cas pour les verbes séparables. Dans les autres cas, nous préférons compter comme en français sur la coïncidence du plus grand nombre possible de mots isolés pour identifier la meilleure phrase-exemple répondant à une question donnée, celle où ces mots reçoivent la traduction particulière due à cette coïncidence.

Finalement, venons-en à l'analyse et à la réduction à ses divers composants lexicaux d'un mot, suite de lettres qui n'est pas séparée par des blancs.

Supposons que l'on nous demande de rechercher le mot *stroom-intensiteitswaarde* = « valeur de l'intensité du courant », et qu'il faille retrouver automatiquement les composantes que le dictionnaire possède à savoir :

stroom	= « courant »
intensiteit	= « intensité »
waarde	= « valeur ».

Nous créons automatiquement à partir de ce mot demandé une série d'hypothèses en coupant tout d'abord tous les préfixes et suffixes appartenant à des listes courtes, et qui auraient pu s'y attacher « librement » aux extrémités.

A l'intérieur des mots composés, la plupart des formes ne sont pas possibles. Ainsi toutes les formes conjuguées d'un verbe ne peuvent pas apparaître à la suture de deux éléments. Nous avons relevé celles qui le peuvent et nous les avons introduites dans une liste, en dérivant toutes les formes admises à partir des unités insérées dans le dictionnaire. Cette liste contient, par exemple, les pluriels aussi bien que les singuliers des noms et la particule /*s*/ qui s'intercalle parfois dans la composition entre deux substantifs.

Comparant cette liste au mot demandé, le programme formule alors toutes les hypothèses de coupures et de non-coupures possibles de ce composé.

Une application récursive de ce processus permet de formuler l'ensemble des hypothèses sur toutes les coupures possibles de chaque mot demandé.

On ne retient parmi ces hypothèses que celles qui sont compatibles avec les règles de composition prévues dans le programme et telles que les fragments obtenus se rejoignent en recouvrant le mot tout entier.

Il est clair que c'est l'hypothèse de moindre découpe qui prévaut; en particulier, parce que le composé peut figurer tout entier comme rubrique dans le dictionnaire, même si ses composants y figurent avec d'autres traductions.

*
* *

Nous dirons pour conclure, que ce dictionnaire, conservé sur cartes perforées et sur bande magnétique, *se présente comme un système documentaire, adapté au traducteur qui l'utilise.*

Il se construit progressivement et son emploi n'est tributaire d'aucune compilation exhaustive de mots ni de racines, si ce n'est de listes très courtes de suffixes et de préfixes « vivants ». Sa mise à jour peut se faire continuellement. On peut aussi en tirer une édition économique et rapide, en offset, ce qui conserve au traducteur le sentiment de sécurité que lui donne le fait de pouvoir feuilleter un vrai volume, — comme il en avait l'habitude, — du moins lorsqu'il ne cherche qu'un petit nombre de termes.

Mais il arrive au traducteur d'avoir un livre entier à traduire dans un certain délai. Le système automatique lui fournit alors une « préparation » du vocabulaire dont il a besoin, *dans l'ordre du texte.* A l'heure actuelle, pour obtenir une telle préparation, le traducteur doit lire son texte, souligner les passages difficiles et les donner à perforer, sans autre précaution.

La lecture automatique des textes pourrait, dans un avenir qui semble relativement proche, lui enlever même cette dernière servitude.