Linguistica Silesiana 32, 2011
ISSN 0208-4228

ALEKSANDRA ŻŁOBIŃSKA-NOWAK
Université de Silesié

SUR LA NOTION DE CLASSE D'OBJETS EN LINGUISTIQUE ET SON UTILITÉ DANS LA DÉSAMBIGUÏSATION DES SENS DES MOTS

ON THE NOTION OF A CLASS OF OBJECTS IN LINGUISTICS AND ITS UTILITY IN THE WORD SENSE DISAMBIGUATION

The main topic of this paper deals with a general description of the notion of the class of objects, its historical draft and development in linguistic studies (generative grammar of N. Chomsky, semantic markers in the theory of J.J. Katz, J.A. Fodor, the *predicate-argument structure* of S. Karolak and the electronic dictionary of G. Gross). The author takes advantage of this concept in her analysis basing on object-oriented approach and the disambiguation by W. Banyś. This article demonstrates also that the disambiguation of the meaning of polysemic word pass by a correct choise of a class of objects which is the only solution to provide adequate equivalents in the target language and the condition of the effective and satisfying translation supported by computer.

Cette présentation portera sur la notion de classe d'objets ainsi que son développement en linguistique. Nous allons tout d'abord proposer une brève esquisse historique de cette notion-là en commençant par le modèle de la grammaire générative de Noam Chomsky et de sa révision sous l'influence des travaux de Jerrold J. Katz, Jerry A. Fodor, à travers la fonction des arguments dans les structures prédicat-arguments de Stanisław Karolak jusqu'à la conception du terme de classe d'objets proprement dite, telle qui a été proposée par Gaston Gross.

Le pas suivant sera consacré à démontrer son importance dans la désambiguïsation des expressions lexicales menant à la traduction automatique effectuée à la base des principes de l'approche orientée objets à la Wiesław Banyś.

Prenant ces données comme base théorique nous apporterons ensuite quelques exemples des classes d'objets servant à la désambiguïsation des emplois choisis des verbes spatiaux français et polonais.

Commençons par la structuration du monde qui s'opère lors du processus de sa connaissance et mène également notre savoir linguistique. Cette activité-là est liée à ce que nous appelons 'ontologie'. Sa définition souligne que c'est une sorte de modélisation de connaissances du monde, d'informations dites extra-linguistiques, organisée en réseau de concepts. Une ontologie s'appuie donc sur des catégories de base (objets, relations, propriétés) qui favorisent la description des objets du domaine traité ainsi que leurs propriétés et les relations entre eux (P. Bouillon, 1998 : 143). Les ontologies portent dans ce sens l'appellation des bases de connaissances ('knowledge base'). En tant que telles elles constituent un modèle idéal qui ne dépend d'aucune langue particulière.

Leur application dans un système de TAL (traitement automatique des langues) est d'une utilité extrêmement importante, à signaler, entre autres : amélioration de la qualité et généralité du système, obtention d'une représentation plus profonde du texte, plus abstraite ou même indépendante de toute langue.

L'ontologie peut aussi fournir certains moyens pour décrypter l'ambiguïté, ce qui ne serait pas possible en basant uniquement sur les données d'ordre lexical, syntaxique, sémantique ou pragmatique, comme en phrase ci-dessous :

Nos voisins ont achevé leurs taureaux et quelques-uns sont tombés.

Si l'on ne se basait que sur des informations lexico-sémantiques on ne serait pas en mesure de déterminer l'antécédent de *quelques-uns*. C'est par une ontologie portant sur la relation de causalité entre deux actions *achever* et *tomber* qu'on arrive à indiquer l'entité qui tombe et à la différencier de celle qui exécute l'action d'achever.

Ce type de procédé rend possible la standardisation du sens par le fait de lever les ambiguïtés de chacun de termes employés dans un domaine de connaissance ainsi que la précision des relations qui existent entre les concepts représentés par les termes, comme celles d'ordre hiérarchique, par exemple, qui donnent l'accès à ce qu'on appelle 'classe d'objets' (est une sorte de, est une instance de, est une propriété de).

Les ontologies ne constituent pas le sujet de cette communication. Cependant leur caractère, les besoins pratiques et les exigences imposées par le TAL permettent aux linguistes de structurer, tout comme dans le cas des classes d'objets, le monde qu'ils observent et qu'ils essaient d'adapter pour le rendre plus intelligible, facile à manier et pour en dégager les informations sur leurs langues.

Passons dans ce contexte à la notion de la classe d'objets. Pour pouvoir en parler il faudrait partir de ses origines et souligner que le phénomène est issu d'un certain type de représentation élaboré dans le domaine de la psychologie. Il est question ici des réseaux sémantiques qui prennent comme point de départ l'idée de l'insertion d'une unité lexicale dans une structure sémantique plus large, le résultat en est que l'unité traitée acquiert du sens par l'intermédiaire de la place qui lui est attribuée dans cette structure-là et des relations qu'elle entretient avec les autres unités de cette même structure.

Les psychologues (cf. p. ex. Collins A. M., Quillian M. R., 1969, 1970) en observant les capacités humaines dans la catégorisation et la mémorisation des concepts ont constaté que la relation la plus évidente et fréquente dans ces deux activités a un caractère hiérarchique et s'opère entre les hyperonymes et les hyponymes comme, par exemple :
autoroute – route – voie de communication – espace à parcourir – espace ou bien *épagneul – chien – animal*, *chêne – arbre – plante* etc. suivant toujours le lien du type 'sorte-de', pour arriver à des relations d'inclusion de classes de différent statut quant à elles-mêmes (les unes superordonnées par rapport aux autres sous ordonnées (infra-ordonnées) aux premières).

Analysons alors le développement de ce phénomène-là en commençant par la description de la langue proposée par Noam Chomsky dans le cadre de sa grammaire générative.

Noam Chomsky s'inspire de l'un des objectifs du M. I. T. (Massachussetts Institute of Technology) qui était la traduction automatique. Il est à remarquer à cette occasion que les nouveaux modèles syntaxiques utilisés en traitement automatique s'appuient sur la version standard (théorie standard) du modèle chomskien et non de son état récent.

Dans la grammaire de Chomsky, qui se veut une conception des structures linguistiques, le rôle prépondérant est accordé à la syntaxe. Comme le souligne l'auteur lui-même, grâce aux phénomènes syntaxiques les utilisateurs de la langue sont capables de construire une infinité de phrases bien formées grammaticalement même si, tel peut en être l'inconvénient, manquant de sens, à rappeler la phrase :

The colorless green ideas sleep furiously.

et sa traduction française *Les idées vertes sans couleur dorment furieusement*.

Le but de cette grammaire est de générer l'ensemble des phrases grammaticales et elles seules et leur assurer une description.

Sont connus dans les travaux de Chomsky les trois types de grammaires, en guise de rappel :

– grammaire à nombre fini d'états considérée comme modèle élémentaire, peu adéquat et performant pour déterminer tous les types de structures syntaxiques ;
– grammaire de constituants se basant sur un système composée d'un axiome de départ, d'un vocabulaire auxiliaire et terminal et d'un ensemble de règles de réécriture ayant pour but d'assurer une dérivation dans un processus qui consiste à produire des phrases ;
– grammaire transformationnelle comportant des règles syntagmatiques servant à engendrer des structures à caractère abstrait et des règles de transformation qui, elles, rendent possible le passage des structures abstraites à de nouvelles structures.

Ce qui reflète dans ce modèle génératif le besoin, plus ou moins naturel, de ranger les unités lexicales dans des groupes bien déterminés et par conséquent, de rendre plus maniable et opératoire les données linguistiques, c'est la construction

du modèle de la grammaire de constituants. Son vocabulaire est divisé en deux sous-ensembles :

– vocabulaire auxiliaire qui permet d'indiquer de grandes catégories grammaticales comme : *GN – groupe nominal, GV – groupe verbal, V – verbe, N – nom,* etc.
– vocabulaire terminal qui classe les unités lexicales, p. ex. *livre, lire, table, fleur* etc. et leur assigne un symbole de catégorie issue du vocabulaire auxiliaire.

Ces deux ensembles font voir clairement que pour construire des structures linguistiques, des phrases comportant plusieurs éléments de nature distincte, il n'est pas suffisant de spécifier des règles qui gèrent leurs combinaisons mais aussi faut-il différencier tous les éléments constitutifs de ces structures-là et préciser leur caractère en indiquant aussi bien des ensembles de nature plus abstraite : *N, V, GN* que plus détaillée : *voiture, orange, amour* etc.

Nous pouvons oser dire que ce que fait Chomsky constitue alors un avènement de la classe d'objets et s'accorde pleinement avec les principes de départ, hérités du groupe M.I.T. dont l'un des objectifs essentiels était, rappelons-le une fois de suite, la traduction automatique.

On ressent alors un fort besoin de classification des éléments linguistiques, une nécessité de trouver de l'ordre et de la hiérarchie dans les constructions produites et tout ceci pour pouvoir appréhender le cheminement des pensées dans les activités langagières et le fonctionnement de la langue en guise de leur reproduction.

Éloigné encore de l'aspect purement sémantique, ces ensembles construits nous offrent une première délimitation de la masse d'informations linguistiques traitée.

Jerrold J. Katz, Jerry A. Fodor inspirés des travaux de Chomsky souhaitaient, quant à eux, associer aux phrases et ses descriptions structurales une ou plusieurs lectures sémantiques pour rendre l'interprétation de ces phrases-là plus renforcée.

Dans l'application d'une composante sémantique ils voyaient également un moyen pour rendre compte de l'ambiguïté sémantique, de l'anomalie ou bien encore de la synonymie présente dans certaines phrases construites correctement à la base des règles grammaticales mais dépourvues de sens. Tout cela devrait concourir à attribuer la (ou les) signification(s) d'une phrase.

Les données sémantiques dans cette nouvelle approche se divisent en deux composantes :

– un dictionnaire qui précise sous forme arborescente les acceptions des lexèmes traités où nous pouvons observer déjà (dans le cas des substantifs) différentes superclasses auxquelles les mots analysés appartiennent, permettant de situer les noms étudiés plus haut dans l'arborescence hiérarchique et étant délimitées par les traits appelés marqueurs sémantiques tels, par exemple : *(non)-matériel, (in)animé, mâle, comestible,* etc.
– un ensemble des règles de projection qui servent à calculer le sens d'une phrase moyennant les informations provenant du dictionnaire et de la description structurale des phrases.

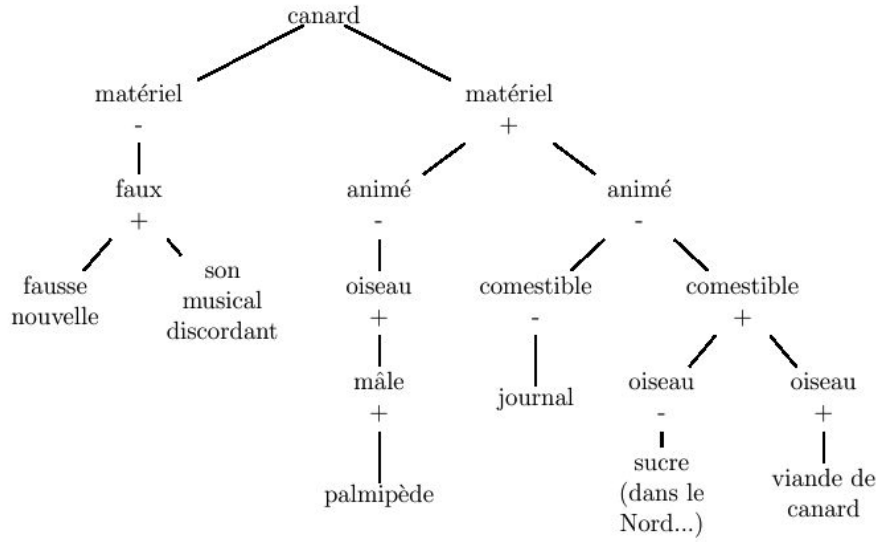

Figure 1 : Analyse sémique de *canard* de Katz et Fodor

En dernier ressort nous pouvons noter que le modèle de la construction arborescente ainsi que le principe de l'inclusion sont ici une solution pour mettre en relief le rôle et la place des unités analysées. Fameux exemple de *bachelor* de Katz et Fodor dévoile davantage que souvent les signifiés des signes sont susceptibles de transgresser les limites catégorielles et en résultat appartenir aux diverses classes sémantiques, ce qui peut, dans cette approche, être résolu par une combinaison de marqueurs sémantiques caractéristique pour chaque signification :

bachelor (célibataire, chevalier, étudiant, phoque) N, N_1, …N_k ;

(I) (physique), (animé), (humain), (masculin), (adulte), (jamais marié) ;

(II) (physique), (animé), (humain), (chevalier servant un seigneur loin de chez lui) ;

(III) (physique), (animé), (humain), (titulaire d'un degré universitaire) ;

(IV) (physique), (animé), (animal), (mâle), (mâle d'un phoque se trouvant privé de femelle pendant la saison des amours).

La description des mots appuyée sur un ensemble de traits – marqueurs sémantiques permet dans des cas pareils de lever l'ambigüité. Un marqueur est donc une unité commune à d'autant plus de termes qu'il est situé près de la racine de l'arbre.

Cette démarche présente donc des qualités importantes quant à l'organisation des significations, les traits caractéristiques des lexèmes et les affinités entre eux. Elle reprend des procédés classiques de la décomposition en „traits pertinents"

tout en étant une proposition nouvelle, frappante, un chemin vers la compréhension, l'interprétation et le traitement des langues naturelles.

D'un côté nous y observons des ensembles – classes très riches, composées de plusieurs éléments – objets comme *animé* qui pourraient être qualifiées de classèmes chez François Rastier et de l'autre, des classes n'en comportant que quelques-uns, par exemple : *viande d'oiseau*.

Néanmoins, même ces petits ensembles peuvent avoir une grande influence et jouer sur la désambiguïsation des lexèmes verbaux dans le cadre de la traduction automatique comme nous le montrerons plus loin.

Bien entendu, une analyse reposant uniquement sur la sémantique lexicale serait insuffisante. Elle doit être accompagnée d'une analyse approfondie des relations syntaxiques si l'on veut décrire le(s) sens d'une phrase. Pour ce faire nous pouvons appliquer en linguistique les représentations en termes de prédicat-arguments. La conception de la grammaire à base sémantique dont l'invention est assignée à Stanisław Karolak (S. Karolak, 1984, 1991, 2007), rend possible la description des propositions à la base des relations prédicatives constituées de prédicats accompagnés d'arguments qu'ils impliquent. Ces structures peuvent être ensuite transformées en phrases quand des unités lexicales concrètes viennent remplir respectivement les positions prédicatives et argumentatives, par exemple :

g (x, y) – monter (Jean, colline) – Jean monte sur une colline.

Nous pouvons distinguer deux types d'arguments impliqués par les prédicats :

– arguments individuels étant des indications individuelles ou des objets physiques ou bien des ensembles d'objets :

Jean (x) *s'est marié* (g) *avec Sylvie* (y). g (x, y)

Elles (x) *ont visité* (g) *toute la Pologne et la France durant leurs vacances d'été* (y).

g (x, y)

– arguments propositionnels qui constituent une proposition cachée ou non derrière la position d'argument

Il (x) *m* (y)*'a informé* (H) *que sa voiture était tombée en panne* (p). H (x, y, p)

Pierre (x) *sent* (G) *la peur* (p). G (x, p)

(*la peur* reflétant toute une situation où quelque chose fait peur à Pierre).

Cette délimitation des arguments favorise la reconnaissance de leurs types. Dans cette optique nous pouvons parler donc des deux classes possibles qui peuvent apparaître à l'entourage des prédicats. Les deux types rendent possible la compréhension du sens des prédicats analysés, une communication exhaustive qui ne prête pas à la confusion entre les locuteurs ainsi que la formation des structures syntaxiquement et sémantiquement correctes.

La grammaire à base sémantique est liée à la traduction automatique des langues par le souci d'expliciter toutes les données nécessaires et de créer un système de règles qui, dans le cas de la TAL, serait susceptible d'être manipulé et transformé par une machine. Cette approche met l'accent sur l'importance des compléments qui se trouvent autour d'un verbe analysé.

C'est donc un formalisme combiné avec des données sémantiques qui peut être appliqué dans la désambiguïsation du sens des mots sans être, lui-même, un outil lexicographique proprement dit.

Gaston Gross dans ses travaux sur le dictionnaire électronique (G. Gross, 1992, 1994a, b, 1995) souligne qu'il faut doter les dispositifs automatiques, servant à générer et reconnaître les phrases d'une langue naturelle, d'informations et d'indications très précises, plus précises et explicites que dans les dictionnaires papier. L'objectif de ce type de description est également de rendre compte de la totalité des emplois d'un mot.

Le problème de la polysémie ne peut être étudié qu'au sein d'une phrase étant une unité minimale de description, son analyse exige aussi la spécification des données de nature syntaxique. Ainsi apparaît-elle comme solution la notion de classe d'objets.

La classe d'objets sert à décrire les ensembles sémantiques homogènes qui possèdent des propriétés syntaxiques spécifiques p.ex. *les vêtements, les moyens de transport*, etc.

C'est le rapport entre l'opérateur d'une phrase et son domaine d'arguments qui gère les relations sémantiques présentes dans cette même phrase. Prenons une forme morphologique *jouer* qui possède plusieurs emplois, nous remarquons facilement que chaque emploi a des domaines d'arguments spécifiques :

Marion joue de la guitare. – toucher

Ton voisin m'a joué. – tromper

Russell Crowe a joué dans la nouvelle version de Robin des bois. – tourner

Chaque emploi du verbe *jouer* implique donc différents arguments. Leur nature peut être notée à l'aide des traits syntactico-sémantiques, comme : *humain, concret, abstrait,* etc.

Même si certains prédicats n'exigent pas de restrictions sur le sémantisme de leurs arguments comme dans les exemples suivants :

Je pense à N.

N me plaît

où *N* peut être réalisé par tout type de nom (ou groupe nominal), on ne peut pas se limiter à l'indication des arguments uniquement par les signes N_0, N_1 etc. puisque les schémas syntaxiques créés ainsi pourraient constituer le premier pas dans la création des phrases aberrantes, par exemple:

jouer – N_0 *jouer* N_1 : * *La table joue la viande*.

* L'amour joue une maison.

Gross propose alors de caractériser ces arguments en ayons recours à des traits sémantiques (reliés au prédicat par des contraintes de sélection) du type : *humain/non humain ; animé/inanimé ; concret/abstrait*, etc. et de surpréciser certaines classes d'objets là où les traits ci-dessus ne sont pas suffisants, par exemple :

N_0 *jouer* N_1 :

N_0 : humain ; N_1 : instrument (pour un des emplois du verbe *jouer*)

Pour mieux fonder ces constations soulignons encore qu'on peut trouver d'autres groupes de prédicats qui choisiront des arguments spécifiques – *locatifs* pour des

verbes spatiaux, par exemple: *aller à, venir de, sortir de, monter à, passer par*, etc.

Les classes d'objets favorisent en conséquence une description souple des substantifs et monosémisent des verbes à de nombreux emplois. Tout ceci ayant pour objectif une traduction aussi convenable que possible des verbes polysémiques.

L'approche orientée objets à la Wiesław Banyś (W. Banyś, 2002, 2005) suit la même lignée de recherche insistant sur le fait qu'un prédicat donné peut être décrit uniquement par le prisme de son emploi. Sous l'emploi d'un prédicat nous comprenons l'indication de son schéma d'arguments définis par les classes d'objets.

En recourant à la notion des classes d'objets on est en mesure de préciser davantage la nature des noms analysés influant sur la compréhension du verbe.

Il est frappant de noter qu'à l'entrée des analyses sur la désambiguïsation des verbes polysémiques le nombre de classes n'est pas connu. Leur nombre dépend entièrement de tous leurs emplois et ne peut pas être précisé d'avance. On procède alors d'abord à la détermination des équivalents résultant de la traduction qui nous permettent ensuite d'en décider.

Il arrive souvent que la spécification des traits sémantiques attribuée à chaque classe construite est trop générale : *humain, animé, concret* etc. Cette caractéristique doit être alors plus détaillée pour assurer une traduction correcte dans la langue cible.

Prenons quelques exemples :

X – [ANM] – ***monter*** – Y – [CONC <objet d'accès en pente>] – **wejść/wchodzić *po***

Ma voisine monte l'escalier avec beaucoup de peine.

Les enfants jouent à monter les marches de l'escalier.

X – [ANM] – ***monter*** – *avec/en/par* – Y – [CONC <appareil mécanique servant à monter>] – **wjechać/wjeżdżać *czymś***

Pour arriver au huitième étage il faut monter avec l'ascenseur / en ascenseur

Je monte par un escalator sur le quai de la voie A.

X – [ANM <animal vivipare ; mâle qui monte sur une femelle>] – ***monter*** (v.tr.) – Y – [ANM <animal vivipare ; femelle sur laquelle monte un mâle>] – **pokryć/pokrywać**

«De vieux chevaux qui n'avaient plus la force de monter la jument sans l'aide du palefrenier» (Buffon).

L'étalon monte la jument.

X – [ANM hum] – ***sortir*** – (*à* – Y – [ANM]) – Z – [CONC <paroles ; traitements outrageux>] – **rzucać/wyrzucać *coś komuś***

Elle sort des injures racistes derrière mon dos.

Après être arrivé en retard, il sort des insultes qu'il nie par la suite.

Nous pouvons donc être amenés dans la désambiguïsation du sens des verbes analysés à la constitution des classes utilisables peut-être une seule fois, applica-

bles à un seul verbe. Ce qui devient essentiel c'est le dressement des listes les plus complètes possibles de chaque classe d'objets qui, en dernière instance, pourrait contribuer à une analyse méthodique de tout le lexique des noms. En d'autres mots, c'est l'indication un par un de tous les membres de chaque classe activée. Certaines classes comporteront plusieurs objets tandis que d'autres seront limitées à quelques-uns. Grâce au système de l'héritage sémantique il sera possible de retrouver les classes dans d'autres classes super ou sous-ordonnées à elles, suivant la règle de l'inclusion, tel est l'objectif de la conception présentée. C'est un travail très minutieux mais en même temps très intéressant et productif.

Cette présentation a été limitée à quelques exemples illustratifs, nous espérons néanmoins que certaines réflexions et observations formulées dans ce travail pourront contribuer à une meilleure compréhension de la notion de classe d'objets ainsi que de son utilité en linguistique contemporaine.

Références citées

Banyś, W. 2002. Bases de données lexicales électroniques – une approche orientée objets : Partie I et II. *Neophilologica 15*: 7–29 et 206–249.

Banyś, W. 2005. Désambiguïsation des sens des mots et représentation lexicale du monde. *Neophilologica 17*: 57–76.

Bouillon, P. 1998. *Traitement automatique des langues naturelles*. Paris : Bruxelles, Éditions Duculot.

Bünting, K.-D. 1989. *Wstęp do lingwistyki*. Warszawa: PWN.

Collin, A. M., M. R. Quillian 1969. Retrieval time for semantic memory. *Journal of Verbal Learning and Verbal Behavior 8*: 240–247.

Collin, A. M., M. R. Quillian 1970. Does category size effect categorization time? *Journal of Verbal Learning and Verbal Behavior 9*: 432–438.

Desclés, J.-P. 1987. Réseaux sémantiques: la nature logique et linguistique des relateurs. *Langages 87*: 55–78.

Delas, D. 1978. La grammaire générative rencontre la figure. Lectures. *Langages 51*: 65–104.

Fuchs, C. 1993. *Linguistique et traitements automatiques des langues*. Paris: Hachette Supérieur.

Fuchs, C., P. Le Goffic 1992. *Les linguistiques contemporaines – repères théoriques*. Paris: Hachette Supérieur.

Gross, G. 1992. Forme d'un dictionnaire électronique. In *La station de traduction de l'an 2000*. Presses de l'Université du Québec.

Gross, G. 1994a. Classes d'objets et synonymie. In *Annales Littéraires de l'Université de Besançon. Série Linguistique et Sémiotique. Vol. 23*: 93–102. Besançon.

Gross, G. 1994b. Classes d'objets et description des verbes. *Langages 115*: 15–31.

Gross, G. 1995. Une sémantique nouvelle pour la traduction automatique. Les classes d'objets. In *La Tribune des industries de la langue et de l'information électronique, Perspectives,* numéro spécial (n^{os} 17–18–19): *Traduction et traduction avec outils, le renouveau pour demain*: 16–19.

Karolak, S. 1984. Składnia wyrażeń predykatywnych. In Z. Topolińska (ed.) *Gramatyka współczesnego języka polskiego. Składnia*, 1–248. Warszawa: PWN.

Karolak, S. 2007. *Składnia francuska o podstawach semantycznych*. Kraków: Collegium Columbinum.

Karolak, S., K. Bogacki 1991. Fondements d'une grammaire à base sémantique. *Lingua e Stile XXVI, 3*: 11–48.

Katz, J. J., J. A. Fodor 1967. Structure d'une théorie sémantique avec applications au français. *Cahiers de Lexicologie n° 9*: 39–72 et *n° 10*: 47–66. Paris: Didier-Larousse.