Krzysztof BOGACKI
Uniwersytet Warszawski
kbogacki@gmail.com

LE MOT, L'ENTITÉ NOMMÉE ET LES DÉFINITIONS STIPULATIVES

1. INTRODUCTION

Dans cet article, nous défendrons la thèse suivante. Le terme de *mot* est sémantiquement flou. Sa référence ne peut être déterminée de façon sûre ni pour une langue donnée ni, à plus forte raison, pour toutes les langues. Ces défauts sont à l'origine de l'émergence de termes et de nouveaux concepts forgés par tous ceux – linguistes, documentalistes, mathématiciens et autres – qui pour leurs recherches ont besoin d'une unité opérationnelle de base. Le mot apparaît souvent dans l'histoire des recherches linguistiques. Le flou foncier dont il semble entaché s'accompagne d'une extrême polyvalence dans son usage, ce qui vient du fait qu'au cours des siècles, ce terme a été employé dans différents contextes et s'appliquait à des réalités linguistiques variées. On mesure les difficultés à trouver une définition descriptive qui cherche un dénominateur commun à tous ces usages. Elles disparaissent dans les définitions stipulatives utilisées lors de la création de termes nouveaux auxquels elles attribuent un sens précis sans tenir compte des emplois traditionnels qui, dans ce cas-là, n'existent pas.

2. LE MOT – UN CONCEPT FLOU

Quoi de plus intuitif que le mot ? A première vue, ce concept semble tellement évident que l'on ne soupçonne même pas les problèmes qu'une

revue rapide des avis sur la façon dont on peut l'appréhender fait surgir. Or, à la réflexion, on s'aperçoit qu'ils sont à tel point importants que l'on finit par se demander s'il ne vaut pas mieux suivre l'exemple de V. Quine qui en écrivant *Le Mot et la Chose* se souciait peu de la définition exacte de ce terme dont le sens restait flou ?[1]

Le *mot* est bien enraciné non seulement dans le discours technique des linguistes mais aussi dans l'usage courant qui peut être saisi à travers les définitions qu'on en donne dans les dictionnaires et encyclopédies. Le sens qu'on attribue à *mot* dépend dans une large mesure de la tradition linguistique locale et de la langue décrite. Au vu de ces deux facteurs, il ne saurait être considéré comme universel.

Commençons par le polonais, où le mot est une unité de la langue qui peut être définie à différents niveaux. Il a ceci de particulier que les référents de ces définitions ne se recouvrent que partiellement :

> « jednostka językowa, definiowana w różnych płaszczyznach, przy czym zakresy wyznaczane przez te definicje tylko częściowo pokrywają się ze sobą » ('unité linguistique, définie à différents niveaux ayant ceci de particulier que les étendues référentielles délimitées par ces définitions ne se recouvrent que partiellement', Encyklopedia PWN).

Il est composé d'un ou de plusieurs morphèmes. En polonais, on distingue parfois *słowo* (« elementarna część mowy » 'unité élémentaire de la langue') et *wyraz* – sa contrepartie graphique (« jego pisanym odpowiednikiem jest wyraz »). Les définitions soulignent le caractère bipartite du mot qui, à ce titre, apparaît comme un signe avec une face signifiante (graphique ou phonique) et une face signifiée. L'attention est attirée sur les limites du mot marquées par les espaces ou les signes de ponctuation :

> « zbiór głosek, który w zapisie graficznym oddzielony jest od innych zbiorów spacjami bądź znakami interpunkcyjnymi » ('ensemble de sons qui, dans la graphie, est séparé d'autres ensembles par des espaces ou par des signes de ponctuation')[2].

[1] Pour sa part, J. Rey-Debove (1971 : 195) observe que les usagers des dictionnaires n'ont aucun besoin « d'une description sémantique totale qui soit juste et précise » et que, adoptant une attitude « pragmatique », ils se contentent souvent d'une approximation sans chercher une définition sans faille.

[2] http://www.alfabet.24on.pl/index.php?title=Wyraz

Les mêmes éléments se retrouvent plus ou moins dans d'autres langues. Ainsi, dans l'usage français, le mot est

> « un élément de la langue composé d'un ou de plusieurs phonèmes, susceptible d'une transcription écrite individualisée et participant au fonctionnement syntactico-sémantique d'un énoncé » (www.larousse.fr/dictionnaires).

Le Petit Robert, quant à lui, distingue le sens courant du mot (« chacun des **sons** ou groupe de sons correspondant à un sens, entre lesquels se distribue le langage ») et le sens linguistique (« forme libre douée de sens qui entre directement dans la production de la phrase »). Le Trésor de la Langue française définit le mot par la formule suivante :

> « Son ou groupe de sons articulés ou figurés graphiquement, constituant une unité porteuse de signification à laquelle est liée, dans une langue donnée, une représentation d'un être, d'un objet, d'un concept, etc. »

L'usage anglophone est reflété par la définition suivante de *word*, donnée dans le Oxford English Dictionary :

> An element or unit of speech, language, etc. Any of the sequences of one or more sounds or morphemes (intuitively recognized by native speakers as) constituting the basic units of meaningful speech used in forming a sentence or utterance in a language (and in most writing systems normally separated by spaces); a lexical unit other than a phrase or affix; an item of vocabulary, a vocable (http://www.oed.com/).
>
> A single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed (Oxford : http://www.oxforddictionaries.com/definition/english/word).

WordReference Random House Learner's Dictionary of American English © 2015 (http://www.wordreference.com/definition/Word) propose :

> a unit of a language, consisting of one or more spoken sounds or their written representation and functioning as a carrier of meaning (WordReference Random House Learner's Dictionary of American English © 2015),

tandis que (http://www.macmillandictionary.com/dictionary/british/word_1) se borne à définir le mot comme « a single unit of written or spoken language ».

Quels sont les éléments les plus importants de ces définitions ? Le mot est considéré, dans l'usage courant, comme une unité de base, intuitivement saisie par les usagers de la langue[3], susceptible de se combiner avec d'autres pour former une unité linguistique d'ordre supérieur. Comme tout signe linguistique, le mot a deux faces, il a ses caractéristiques propres aux deux niveaux : phonétique et graphique, correspondant à des réalités différentes selon différentes langues.

Vu par les non-linguistes, le mot est considéré souvent, surtout en littérature, comme l'unité minimale de l'analyse. Parmi les linguistes, les avis sont partagés ; d'autres candidats au rôle d'unité de base apparaissent, dont le morphème. Il serait cependant erroné de prétendre que le mot, en usage en linguistique depuis fort longtemps, est oublié. Il a toujours sa place même dans les théories récentes, où il est adopté comme unité primaire en dépit de ses bases théoriques fort douteuses.

Toute la tradition grammaticale est fondée sur la notion de mot et les premières réflexions remontent à la tradition grammaticale en Inde. Distingué chez Panini de la notion plus générale de *parole* ou d'*expression langagière*, le concept grammatical de mot s'y oppose à la phrase et au phonème.[4] La tradition européenne – grecque et latine – fait également appel à cette notion. Jusqu'à Platon, le mot était confondu avec le nom. A partir de la période hellénistique, concurrencé par le terme *lexis*, le *mot* est appréhendé comme une 'partie du discours' (Lallot 1992 : 125). Dans *De lingua latina*, Varron, de son côté, adopte le mot comme point de départ de ses analyses sans s'embarrasser des difficultés théoriques à le définir de façon précise. Même de nos jours, les auteurs de grammaires tout en faisant appel à la notion de morphème, l'utilisent à titre de concept théorique sans l'appliquer à des analyses concrètes, et surtout sans en faire l'élément central organisant la description d'une langue. Touratier (2002) déplore cette situation rappelant que dans l'usage grammatical, c'est le mot qui « constitue incontestablement le signe de base » (Riegel, Pellat, Rioul 1994 : 558) et non pas le morphème qui contrai-

[3] Souvent les intuitions des sujets parlants moyens ne concordent pas avec la vision des mots des experts. On en trouve des exemples dans des copies d'élèves et même dans la presse : *aparça* (= *à part ça*), *jepar anvacans* (= *je pars en vacances*), *masoeur* (= *ma soeur*), *vendredisoir* (= *vendredi soir*). Cf. M.-J. Reichler-Béguelin (1988).

[4] Cf. G.-L. Pinault (1992). Dans la réflexion moderne, il se situerait plutôt entre le monème (ou le morphème) et le syntagme.

rement au mot est un concept parfaitement définissable et rigoureusement défini.

Le terme de mot a plusieurs défauts qui devraient lui faire perdre la place importante qu'il garde toujours dans les analyses linguistiques. Il n'est ni aisément définissable ni facilement délimitable. En linguistique structurale, il est évité à cause de son manque de rigueur (Dubois et al. 1973 : 327). Certains le rejettent, en lui substituant le morphème ou le monème (entendu comme l'unité significative minimale dans le système martinétien).[5] De nos jours, c'est « le morphème [qui] désigne le plus petit élément significatif individualisé dans un énoncé, que l'on ne peut diviser en unités plus petites sans passer au niveau phonologique » (Dubois et al. 1973 : 324). La difficulté à définir le mot augmente proportionnellement au nombre de langues dont on tient compte. On s'aperçoit en effet que la définition du mot tenant compte de la réalité langagière devrait varier selon les langues (Vendryès 1968 : 106) allant jusqu'à rendre impossible la délimitation des mots sur le plan graphique comme par exemple en thaï[6].

3. DÉFINIR LE MOT

On s'accorde volontiers pour dire qu'un mot est un signe linguistique qui possède une certaine autonomie de fonctionnement et une certaine cohésion interne. Les difficultés apparaissent lorsqu'on tâche de préciser cette définition trop vague. Les critères distinctifs du mot-signe linguistique s'insèrent soit dans la conception saussurienne (entité à deux faces : signifiant + signifié), soit dans celle de Peirce qui tient compte d'un troisième facteur qui est le référent.[7] Commençons par les caractéristiques formelles.

---

[5] A. Martinet (1965 : 84) déclare carrément que le terme de mot « est inutilisable, aussi bien dans une recherche syntaxique sérieuse que dans la présentation de ses résultats ».

[6] J. Lallot (1992 : 125) note qu'en Grèce ancienne, « sauf dans les écritures syllabiques mycénienne et chypriote, le mot n'a pas d'individualité graphique ». Ce cas n'est pas isolé. On le relève en thaï à cause de son alphasyllabaire particulier et dans d'autres langues orientales.

[7] Le souci de définir le mot avec précision ne se retrouve pas dans tous les grands courants linguistiques. Fréquent dans le courant d'obédience structuraliste, quasiment indispensable en morphologie et en lexicologie, il laisse indifférents les partisans de la grammaire à base sémantique où le concept de mot se trouve à l'arrière-plan. L'idée

3.1. DÉFINITIONS FORMELLES

Dans certaines définitions, on cherche avant tout un **séparateur** permettant de décider où un mot commence et où il se termine. Autrement dit, ce qui est recherché, ce sont les limites du mot. Ainsi le mot serait-il un ensemble graphique sans espaces internes ou, ce qui revient au même, « le mot correspond à une séquence de lettres délimitée par deux blancs : l'un au début, l'autre à la fin ». Il est vite apparu que cette définition était insatisfaisante : l'espace ne peut pas être considéré comme un bon séparateur de mots. En effet, ce critère semble peu sûr dans la mesure où on hésiterait à voir dans *parce que* deux mots. La pratique lexicographique et grammaticale pluriséculaire ferait opter en faveur du traitement unitaire, de même que dans le cas de certains composés tels que *pomme de terre* ou *machine à laver*. Au contraire, dans les articles contractés, on voit parfois deux mots et non un seul en dépit de l'absence d'espace.

D'autres signes formels sont tout aussi peu fiables. Il en est ainsi de l'apostrophe. La séquence sans espace à l'intérieur mais comportant une **apostrophe** telle que *l'enfant* ou *l'énigme* est considérée comme comportant deux mots. Cependant, l'apostrophe ne peut pas être considérée comme une marque de délimitation d'un mot : *aujourd'hui* qui en comporte une est considéré comme un seul mot bien que, d'un point de vue historique, il s'agisse d'une expression polylexicale.

Le même cas de figure se retrouve avec le **trait d'union** : dans *wagon-lits*, il ne divise pas la séquence de lettres en deux mots contrairement à *vient-il* considéré comme bel et bien composé de deux mots. Son statut de séparateur formel, même si l'on décide de lui en accorder un, n'est pas toujours le même. Dans *a-t-il*, on trouvera deux mots tout comme dans *aurait-il* : or dans le premier cas, on trouve deux occurrences de trait d'union, alors que dans le second une seule.

Le **point** ne saurait être considéré, non plus, comme un séparateur fiable. Apparaissant parfois dans les sigles (*I.B.M.* à côté de *IBM*), a-t-il le même rôle que lorsqu'il est employé dans les dates (12.05.2012) ou terminant un titre (*M. Dupont* – pour *Monsieur Dupont*) ou un prénom abrégé

centrale est celle de prédicats sémantiques combinés dans des représentations sémantiques avec leurs restrictions diverses et correspondant à des réalisations de surface variées. L'accent est déplacé vers les propriétés conceptuelles des prédicats sémantiques et des conséquences qui en découlent visibles au niveau de la surface.

(*M. Dupont* pour *Michel Dupont*) ? De même, les symboles digitaux qu'on trouve parfois dans les textes ne suffisent pas pour scinder la séquence de lettres en deux (cf. H_2O – mot simple). De façon générale, on devrait rappeler que l'usage des marques formelles – espaces, apostrophe, trait d'union et point – est ambigu et reflète plutôt des habitudes changeantes fixées dans la langue au cours des siècles que la réalité qu'on aimerait trouver dans la langue. Il serait sans doute opportun de rappeler ici que l'orthographe actuelle reflète les tendances, les erreurs commises par les grammairiens, et les malentendus de toute sorte accumulés tout au long des siècles sans égard au sentiment des usagers de la langue capables d'identifier les mots. Se sont ajoutées à cela les habitudes des imprimeurs qui, par leurs choix en matière d'orthographe, ont fini par imposer certains usages. A partir du XVI^{e} siècle, on assiste en France à des tentatives plus ou moins réussies de réforme de l'orthographe. Les controverses entre « étymologistes » et « phonétistes » n'ont pas toujours conduit à des résultats où le bon sens serait sauf. Il n'était pas rare de voir deux graphies différentes pour une seule et même expression : l'une conduisant à une interprétation unitaire, l'autre au contraire à une interprétation en deux mots. Ainsi, on relève chez Beauzée (1767 : vii) :

> « Je donne, à cette succession, le nom d'ordre analytique **parce qu**'elle est tout à la fois le résultat de l'analyse de la pensée, & le fondement de l'analyse du discours dans toutes les langues. »

Vingt années plus tôt, l'abbé Girard (1747 : 65–66), au contraire, écrivit :

> « quoique vous m'ayez enseigné la vertu, j'ai préféré la débauche ; **parce que** la nature m'a donné des inclinations plus fortes que vos conseils. »

En ce qui concerne l'usage actuel, on évoquera une recherche de M. Mathieu-Colas (1988) qui établit un inventaire de mots composés avec une graphie différente. Il relève de nombreux désaccords du point de vue orthographique entre le Petit Larousse et le Petit Robert et un certain nombre de contradictions internes. Les hésitations concernent entre autres l'utilisation du trait d'union (p. ex. *audio-visuel* ou *audiovisuel*, *fille-mère* ou *fillemère, portemanteau* ou *porte-manteau*). Il serait absurde, à notre avis, de maintenir, à la faveur de l'orthographe changeante, que ces mots-là sont tantôt des mots simples tantôt des séquences de mots.

Pour revenir à la liste des candidats séparateurs formels, on ajoutera un facteur supplémentaire : c'est l'accent. En polonais, il se place

sur la pénultième. Comme pour les autres critères, il est facile de trouver des contrexemples de mots accentués sur la troisième syllabe à compter de la fin : *republika*, *pedagogika*, *muzyka*, *matematyka*, *zgadywaliśmy* etc. et même sur la quatrième (*robilibyśmy*, *zgadywalibyśmy*, *pisalibyśmy* etc.). Il est plus sûr en français, où l'accent sur la finale ou sur la pénultième dépend entièrement de la structure de la syllabe et où ces deux situations sont les seules à observer. Dans les deux langues, la situation est brouillée dans le cas des groupes rythmiques formés par accumulation de plusieurs formes.

D'autres critères formels qu'on invoque de temps en temps (Giurescu 1975 : 29) échouent également, laissant échapper les prépositions et les clitiques – p. ex. celui du comportement morphosyntaxique unitaire se traduisant par l'autonomie syntaxique (le mot serait la plus petite unité « libre », susceptible de faire énoncé à elle seule) ou flexionnelle (la plus petite unité avec un flexif).

Force est de constater que les critères invoqués jusqu'ici – pris isolément ou combinés entre eux – ne permettent pas de forger une définition du mot qui ne laisse pas de résidu. De plus, en ce qui concerne les traits graphiques, ils suscitent une critique de taille, qui consiste à dire que mettre l'accent sur la forme écrite du langage revient à présupposer l'importance foncière des divisions du texte écrit, ce qui va à l'encontre de l'opinion généralement admise que la nature réelle du langage humain ne peut être saisie qu'à sa manifestation phonique. Qui plus est, les tentatives de définition du mot que nous venons de voir obnubilent le mot en tant que signe linguistique, c'est-à-dire que le mot est une unité à deux faces : la face signifiante et la face signifiée. S'attachant au côté formel, elles ignorent totalement le côté sémantique et, si on est partisan de la conception peircienne, le côté référentiel.

3.2. CRITÈRES SÉMANTIQUES

Nous venons de voir que les mots ne sont pas reconnaissables par leur forme. Le seraient-ils par leur sens ? Un mot pourrait-il être caractérisé comme la plus petite unité de sens possible ? On éviterait ainsi le reproche de la primauté sous-entendue de l'aspect formel et de l'élimination de l'autre face – signifiée – du mot.[8] Là encore, le critère se révèle insatis-

[8] Cf. la définition de J. Lallot (1992 : 125), qui, s'appuyant sur une tradition de la

faisant et entre parfois en contradiction avec le critère formel de l'espace séparateur. Ainsi, il devient embarrassant dans le cas de l'allemand, où les composés se construisent par attachement d'éléments constitutifs faisant disparaître les espaces qui existent bel et bien en français (cf. *Datenbearbeitung* vs. *traitement des données*). D'autre part, si l'on élimine de la définition tout critère sémantique, on tombe dans l'excès contraire, qui consiste à négliger le côté signifié du mot, qui constitue cependant un de deux facteurs définitoires du signe linguistique.

Résumons-nous. Tout essai de définir le mot – soit à l'aide d'un critère unique, soit avec une combinaison de critères variés – conduit à l'échec. Il s'avère notamment que les critères utilisés dans les définitions ne permettent pas de couvrir tout ce qui est désigné par ce terme dans la pratique courante et qu'il est difficile voire impossible de faire correspondre définition et réalité. L'impression qui se dégage de la confrontation de la réalité linguistique désignée par ce terme est qu'avec le mot on entre dans le domaine du vague et de l'imprécis, où les frontières sont floues et les points de référence incertains. En effet, à côté des cas où le terme de mot s'applique clairement et recouvre ce que la pratique courante considère comme mot, on relève « des cas où il est clair que le terme ne peut pas s'appliquer » et « des cas où le locuteur ne peut décider si, étant donné le sens que possède l'expression, celle-ci s'applique ou non (sans que cette impossibilité soit due à un manque de connaissances de la part du locuteur) ».[9]

Dans les définitions du mot que nous venons de mentionner, il existe, certes, un noyau dur [suite de sons/séquences de caractères] composé d'un nombre limité d'éléments définitoires, mais il conduit à une impasse : cette tentative de cerner le mot laisse un résidu considérable.

En dépit d'échecs renouvelés, on revient à la charge en essayant de trouver une formule définitoire satisfaisante. Mentionnons, à titre d'exemple, deux colloques qui ont réuni les représentants de disciplines diverses et qui dans leur panoplie trouvent ce terme : celui sur « Le Mot en traduction automatique et en linguistique appliquée »

grammaire grecque, combine les critères relevant de deux faces. Il définit notamment le mot comme « un segment compact, porteur d'un accent unique et d'une signification non composée ». Ces critères paraissent toutefois peu clairs.

[9] Cf. http://www.semantique-gdr.net/dico/index.php/Ambigu%C3%AFt%C3%A9 consulté le 09 juillet 2015.

qui s'est tenu le 8 décembre 1962 suivi de la publication d'un recueil d'interventions, ainsi que celui de novembre 1988[10]. Ni l'un ni l'autre n'ont permis d'arriver à une formule unique qui mettrait en facteur différentes approches et concilierait les préoccupations divergentes: celles des mathématiciens, sémioticiens, informaticiens et linguistes. Même à l'intérieur d'une seule branche du savoir, des divergences d'intérêt se faisaient clairement sentir. Ainsi, parmi les linguistes, B. Pottier considérait le mot comme « une unité de comportement » tandis que Cl. Dubois, en bon lexicographe, voyait dans le mot *l'adresse*, « l'unité délimitée par deux blancs typographiques réduite à la forme de paradigme considérée comme fondamentale ». Ce qui frappe et ce qui, en un sens, est normal, c'est que chacune des approches correspond aux **besoins** des spécialistes des domaines divers et reflètent les particularités de la matière linguistique traitée. Ainsi, les lexicographes dans leur pratique quotidienne font face à des entrées de dictionnaire simples délimitées par des espaces, ou au contraire composées de segments avec un ou plusieurs espaces à l'intérieur, sans ou avec trait d'union, sans ou avec apostrophe. D'un autre côté, l'analyse morphologique mettant en scène les catégories grammaticales telles que le cas, le genre et le nombre à propos des formes telles que *rosarum*, *dominis*, etc. conduit aux notions de monèmes et de synthèmes. Il semblerait qu'une mise en facteur commun des définitions avancées soit impossible. En effet, elle conduirait à une formulation qui ne resterait pas exempte de contradictions à un niveau de généralité la rendant non-opératoire dans des situations précises et la privant de sa qualité recherchée au départ : celle de l'utilité pratique.

Nous pensons que la question de savoir comment définir le mot avec, comme arrière-pensée, l'espoir d'arriver à une réponse unique acceptable pour les chercheurs venus d'horizons divers, est vaine. La question : 'qu'est-ce que le mot ?' doit être replacée dans la perspective de l'utilisateur particulier par la question 'qu'est-ce que le mot pour X ?'.

4. LE MOT ET SES COUSINS

Au vu de ce que nous venons de rappeler ci-dessus, le mot apparaît comme une réalité polyvalente susceptible de s'adapter à des usages mul-

[10] Colloque de Paris IV sur le Mot, novembre 1988. Cf. Fruyt & Béguelin (1990).

tiples pour répondre à des besoins variés. Les difficultés de tous ordres qu'on rencontre en voulant donner une définition rigoureuse du mot ont fait naître toute une série de concepts gravitant autour de ce terme mal défini et pourtant très utile. Les considérations théoriques sur le mot conduisent aussi vers des notions voisines telles que lexème, lexie, synapsie, morphe, morphème, monème, synthème, syntagme, phrasème, composé, pragmatème, idiome, proverbe, etc.

A partir des années 1980, un concept nouveau, proche par le nombre de ses traits de celui du mot, est venu s'ajouter à cette liste. Il s'agit de l'entité nommée (EN), qui a émergé petit à petit, stimulée par le développement spectaculaire du TAL. Les circonstances de sa naissance prouvent, une fois de plus, que les nouveaux concepts qui apparaissent sont autant de réponses à des besoins précis créés sous l'impulsion de conditions changeantes et de facteurs externes à la langue. Nous pensons en particulier à l'observation suivante maintes fois formulée par les informaticiens : la fouille de données est plus efficace, le repérage dans un texte des concepts clés et l'extraction d'éléments pertinents d'information (pour identifier, par exemple, des occurrences d'événements particuliers) se fait plus rapidement si l'on dispose d'entités dont la nature a été spécifiée préalablement. Le chemin était ainsi ouvert pour l'avènement de 'l'entité nommée'. Elle se trouve ainsi au coeur des problèmes du développement des systèmes automatiques d'indexation et d'analyse du contenu de documents comme la veille, les résumés automatiques et d'autres. Comment pourrait-on préciser sa nature ?

5. L'ENTITÉ NOMMÉE OU LA NOUVEAU-NÉE COUSINE DU NOM PROPRE

Cette tâche a trouvé un cadre de développement grâce à la série de 7 conférences MUC (Message Understanding Conferences), organisées par diverses institutions américaines, qui se sont déroulées de 1987 à 1998. C'était en réalité des campagnes d'évaluation des systèmes servant à extraire des informations précises. La formule commune à toutes ces conférences était la suivante : après avoir précisé le type d'information recherché à extraire d'un texte-cible, on distribuait aux équipes des textes d'apprentissage pour évaluer à la fin les résultats obtenus. Les résultats pouvant être exploités à des fins militaires (contribuant à accélérer

la compréhension des messages militaires), on comprend que les travaux aient été financés par la DARPA (Defense Advanced Research Projects Agency). Il y a eu trois 'cycles' de conférences, la première en 1987, la dernière en 1993. Dans un premier temps, c'est vers l'anglais que les recherches se sont orientées. Assez rapidement, d'autres langues se sont trouvées sur le chantier : l'espagnol, le hollandais, le chinois, l'allemand et d'autres[11].

Conformément aux exigences formulées tout au début, on recherchait trois types d'EN :

– celles qui correspondent à des noms de personnes, d'organisations et de lieux. Elles formeraient la classe appelée ENAMEX et seraient subdivisées en trois sous-types : PERSON, ORGANISATION et LOCATION ;
– les expressions temporelles, qui appartiendraient à la classe TIMEX avec deux sous-types : DATE et TIME ;
– les expressions numériques, de monnaie et de pourcentage, qui formeraient la classe NUMEX avec deux sous-types : MONEY et PERCENT.[12]

L'étendue du champ recouvert par l'EN est restée limitée même si, à la suite des travaux ultérieurs entrepris dans différents centres de recherches, on a vu la liste de sous-types s'allonger considérablement en fonction de la granularité descriptive. Ainsi, il est possible d'ajouter au groupe de NUMEX les adresses postales et électroniques, les numéros de fax et de téléphone. W. Paik et al. (1996) tiennent compte de 30 catégories réparties en 9 classes : GEOGRAPHIQUE, AFFILIATION, ORGANISATION, HUMAIN, DOCUMENT, EQUIPEMENT, SCIENTIFIQUE, TEMPORELLE et DIVERS[13], etc.

[11] Cf. M. Ehrmann (2008 : 21), qui mentionne aussi des applications directes en TAL des systèmes de reconnaissance des entités nommées.

[12] Cf. M. Ehrmann (2008 : 17). Les auteurs des conventions adoptées pour ESTER2 (= **E**valuation des **S**ystèmes de **T**ranscription enrichie d'**E**missions **R**adiophoniques) notent que « les informations « temps » et « montant » ne sont pas des entités nommées mais qu'elles sont visées par les tâches d'extraction d'informations. L'ensemble peut être appelé entités spécifiques. » http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

[13] Le nombre de catégories prises en compte par Sekine et al. (2002), Sekine et Nobata (2004) atteint 200.

5.1. DÉFINIR L'EN

Ehrmann (2008) en annexe B donne un inventaire de propos définitoires plus ou moins détaillés recueillis dans des rapports de campagnes d'évaluation et de projets, ou encore dans d'autres sources. Les propositions sont compatibles avec la théorie de Peirce plutôt qu'avec celle de Saussure. Cette liste, établie à peu près une vingtaine d'années après les premières discussions sur les EN est sans doute incomplète. S'il est un point sur lequel tout le monde est d'accord, c'est pour dire qu'il faut écarter la formule définitoire énumérative. Rares sont d'autre part les définitions qui rapportent l'EN à un modèle informatique comme dans :

> « Une entité nommée est une expression linguistique autonome. On dit d'une expression linguistique qu'elle est autonome référentiellement quand elle peut, par ses seules ressources, évoquer un référent. [...] Etant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »[14]

Contrairement au mot, qui englobe aussi bien les formes sémantiquement pleines que les formes vides, à valeur grammaticale, les EN ne visent que les unités et les structures pourvues de la capacité de référer. On s'accorde à dire que l'EN est discursivement **monoréférentielle** contrairement au nom propre polyréférentiel[15]. « Rien n'est entité nommée par 'nature', seulement des unités linguistiques monoréférentielles peuvent le devenir, et ce dans le cadre d'une modélisation applicative uniquement. » (Ehrmann 2008 : 167)[16]. Elle est liée à son référent unique grâce à une relation de désignation[17]. Si la plupart des définitions voient comme des EN les noms propres ou les acronymes, la version du 08 décembre 2005 de l'Atalapédia y ajoute « les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologie) » et souligne : « Ces séquences référentielles sont

---

[14] Ehrmann (2008 : 180).

[15] « L'entité nommée est la notion utilisée en TAL pour désigner les eléments discursifs monoréférentiels qui coïncident en partie avec les noms propres et qui suivent des patrons syntaxiques déterminés » (Rangel-Vicente 2005).

[16] C'est cette propriété-là qui permet de réunir dans un même ensemble des choses aussi disparates que les véritables noms propres et les indications de temps et de lieu comme on le fait depuis les premières conférences MUC.

[17] Cf. Kleiber (2001 : 24).

primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes » (Ehrmann 2008 : 257–258).

5.2. DÉFINITION DESCRIPTIVE VS DÉFINITION STIPULATIVE

Il serait opportun, à notre avis, de rappeler un point qui fait la différence entre la définition de l'EN et celle du mot. Avec le *mot*, on se trouve devant un terme et un concept dont on cherche à délimiter l'étendue par une définition de type descriptif, c'est-à-dire visant à rendre compte des usages du terme établis depuis des millénaires et appliqués à différentes réalités langagières. Avec l'EN, au contraire, on est dans une situation de création de concept dont on délimite les contours avec une définition de type stipulatif (ou prescriptif). Celle-ci n'est aucunement contrainte par une tradition d'emploi, elle associe un terme avec une dénotation et un sens précis sans avoir à tenir compte d'une tradition d'usage pluriséculaire. Elle circonscrit le concept de façon à l'adapter au mieux à des besoins précis qui sont à l'origine de la création du terme introduit. Les définitions stipulatives échappent donc à la dichotomie 'vrai/faux', 'trop large/trop étroite' ou 'exacte/inexacte', et peuvent être évaluées comme 'bien ou mal formées' par rapport aux buts qui sont à l'origine de leur création. Une telle situation ne veut pas dire qu'il y ait une et une seule façon de définir l'EN. Bien au contraire, la dépendance d'un besoin précis qui conditionne le sens donné à l'EN fait qu'on note d'ores et déjà des divergences entre les définitions proposées de ce concept.[18] A priori, une définition stipulative réduisant les EN aux noms d'organisations à l'exclusion des dates, des lieux, etc. est tout à fait concevable. D'un autre côté, une recherche ayant pour but de repérer dans un corpus les hydronymes (ou encore les gènes, les maladies, les numéros de fax, etc.) serait ciblée sur ce type particulier d'EN à l'exclusion des autres. Certes, l'extraction d'information via les EN étant un domaine en plein développement, on voit s'allonger chaque jour la liste de termes avec un halo de propriétés – syntaxiques, morphologiques, ou autres – que l'on essaie de dépister dans les corpus. Mais si l'on essaie d'en tenir compte pour une définition du concept d'EN, on risque de tomber dans le

[18] Le Meur et al. (2004 : 4) constatent carrément qu'« il n'existe pas de définition standard ».

piège où tombent ceux qui cherchent désespérément à définir le mot par une formule qui n'est plus stipulative mais descriptive avec les difficultés qui lui sont inhérentes et qui, à un moment donné, risquent de devenir insurmontables.

Evoquons à titre d'exemple la proposition de K. Fort et al. (2009), qui, parmi les critères définitoires des EN, citent la stabilité dénominative tout en y incluant les pronoms personnels (p. ex. *il*). Or, les deux critères s'excluent mutuellement. En effet, on ne saurait admettre que les pronoms personnels se caractérisent par une quelconque stabilité dénominative tout en maintenant, à la suite de Kleiber (2001 : 24), qu'« il n'y a en effet relation de dénomination entre *X* et *x* que si et seulement s'il y a eu un *acte de dénomination* préalable, c'est-à-dire l'instauration d'un lien référentiel ou d'une fixation référentielle, qui peut être le résultat d'un acte de dénomination effectif ou seulement celui d'une habitude associative, entre l'élément *x* et l'expression linguistique *X* ».

6. REPÉRER LES EN DANS UN TEXTE

Quelles sont les techniques permettant de repérer les EN dans un texte ? Mentionnons-en deux. La première, répandue surtout dans les années 90, est appelée symbolique et a un fondement linguistique. Elle est coûteuse car elle demande énormément de ressources humaines et un travail manuel pour écrire les règles. Elle peut être de différents types : morphologique, syntaxique, sémantique, voire pragmatique. Elle fait appel à l'intuition humaine et implique la construction manuelle des modèles d'analyse des textes comportant des descriptions d'enchaînements possibles de syntagmes nominaux ou verbaux correspondant à l'information recherchée, c'est-à-dire constituant des entités nommées. Admettant que *pendant deux ans* constitue une EN de type temporel, on construira dans un formalisme approprié une description de ce type de syntagme disant, par exemple, quelle préposition est admise (*depuis*, *en*, *après*, etc. contrairement à *selon*, *vis-à-vis de*, etc.) et quel est son régime potentiel.

L'autre type d'approche, basée sur l'apprentissage, devenue à la mode ces derniers temps, fait appel à des méthodes statistiques appliquées à de grands volumes de données. Elle consiste à observer des corpus annotés. Ayant noté à plusieurs reprises la présence du mot *docteur* et ayant relevé des noms annotés comme noms de personne, le système conclut à la valeur de l'ensemble : *Docteur X* qui sera considéré

comme une EN de type ENAMEX. Cette approche est plus rapide que la précédente, le coût de développement des systèmes est moins important mais il serait erroné de prétendre qu'elle se passe totalement de l'intervention humaine. Elle a lieu au moment de l'annotation de corpus nécessaires pour l'apprentissage[19].

7. CONCLUSION

Si nous avons mentionné les EN, ce n'est pas parce qu'elles sont susceptibles d'éliminer le concept de *mot*. Ce terme et la réalité à laquelle il réfère ont été cités ici parmi d'autres qui apparaissent en réponse à une sollicitation nouvelle dans un contexte en mutation constante. L'EN telle qu'elle apparaît aujourd'hui laisse ouvertement de côté un grand pan de ce qu'on désigne par le terme de *mot*. Elle a cependant pour elle le mérite de répondre au besoin qui l'a fait naître et se justifie par un taux de réussite en termes de précision extraordinairement élevé. En effet, à l'issue de MUC6, les résultats notés pour les logiciels de filtrage d'EN se sont avérés proches des performances humaines, ce qui a permis la commercialisation de deux systèmes et l'intégration d'autres dans des systèmes gouvernementaux d'analyse de textes[20].

Le risque est évident que l'EN devienne un terme aux emplois multiples, comme le mot et ses cousins qui l'avaient précédée. On ne s'étonnera donc pas de voir de nouvelles acceptions apparaître, créées par des besoins variés, auxquels il sera difficile de trouver un dénominateur commun. Mais on glissera alors vers une définition descriptive de l'EN.

RÉFÉRENCES

Beauzée N., 1767, *Grammaire générale, ou exposition raisonnée des éléments nécessaires du langage, pour servir de fondement à l'étude de toutes les langues*, Barbout, Paris.

Dubois J., Giacomo M., Guespin L., Marcellesi Ch., Macellesi J.-B., Mevel J.-P., 1973, *Dictionnaire de linguistique*, Larousse, Paris.

[19] Il est possible de coupler les deux techniques dans une approche hybride qui semble assez prometteuse.

[20] Cf. M. Ehrmann (2008 : 18).

Ehrmann M., 2008, *Les entités nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Thèse de doctorat, Université de Paris XIII.

ESTER2, 2007, *Entités nommées, dates, heures et montants. Convention d'annotation*, version 0.1. On-line : www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

Fort K., Ehrmann M., Nazarenko A., 2009, *Vers une méthodologie d'annotation des entités nommées en corpus ? Traitement Automatique des Langues Naturelles*, Jun 2009, Senlis, France. On-line : hal.archives-ouvertes.fr/hal-00402321.

Fruyt M., Béguelin M. J., 1990, « La notion de 'mot' en latin et dans d'autres langues indoeuropéennes anciennes », in : *Actes du Colloque de Paris IV sur le Mot, novembre 1988*, 21–46.

Girard G., 1747, *Les vrais principes de la langue françoise ou la parole réduite conformément aux lois de l'usage* (Reprod. en fac-sim.)/abbé Gabriel Girard ; introduction par P. Swiggers, Droz, Genève.

Giurescu A., 1975, *Les mots composés dans les langues romanes*, De Gruyter Mouton, La Haye.

Kleiber G., 2001, « Remarques sur la dénomination », *Cahiers de praxématique* [En ligne], 36/2001, consulté le 10 juillet 2015. URL : http://praxematique.revues.org/292.

Lallot J., 1992, « Le mot dans la tradition prégrammaticale et grammaticale en Grèce », *LALIES* 10, 125–134.

Le Meur et al. 2004, Le Meur C., Gallinao S., et Geoffrois E., *Conventions d'annotations en Entités Nommées*, http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/convention_en_old.pdf.

Martinet A., 1965, « De la morphonologie », *La linguistique* 1, 15–30.

Mathieu-Colas M., 1988, « Variations graphiques des mots composés dans le *Petit Larousse* et le *Petit Robert* », *Linguisticae Investigationes*, 12:2, John Benjamins B.V., Amsterdam, 235–280.

Paik W., E. Liddy, Yu, E. et McKenna, M., 1996, *Categorizing and standardizing proper nouns for efficient information retrieval. Corpus processing for lexical acquisition*, B. Boguraev & J. Pustejovsky (éd.), MIT Press Cambridge, MA, USA, 61–73.

Pinault G.-J., 1992, « Le mot et l'analyse morphologique selon la grammaire indienne », *LALIES* 10, 159–176.

Quine W. V., 1960, *Word and Object* [= *Le mot et la chose*, trad. française par J. Dopp et P. Gochet, Paris, Flammarion, 1978].

Rangel-Vicente M., 2005, « La glose comme outil de désambiguïsation référentielle des noms propres purs », *Corela* [En ligne], consulté le 08 février 2015. URL : http://corela.revues.org/1212

Rey-Debove J., 1971, *Étude linguistique et sémiotique des dictionnaires français contemporains*, Mouton, The Hague, Paris.

Reichler-Béguelin M.-J., 1988, « Perception du mot graphique dans quelques systèmes syllabiques et alphabétiques », *LALIES* 10, 143–178.

Riegel M., Pellat J.-Ch., Rioul R., 1994, *Grammaire méthodique du français*, Paris, PUF.

Sctrick R., 2015, « MOT », *Encyclopœdia Universalis* [en ligne], consulté le 4 janvier 2015. URL : http://www.universalis.fr/encyclopedie/mot/.

Sctrick R., 2015, « SYNTAGME », *Encyclopœdia Universalis* [en ligne], consulté le 11 février 2015. URL : www.universalis.fr/encyclopedie/syntagme/.

Sekine S., Nobata C., 2004, « Definition, Dictionary and Tagger for Extended Named Entities », in : *Forth International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, Portugal, 1977–1980.

Sekine S., Sudo K., Nobata C., 2002, « Extended named entity hierarchy », in : *The Third International Conference on Language Resources and Evaluation (LREC)*, Iles Canaries, Espagne.

Touratier Ch., 2002, *Morphologie et morphématique. Analyse en morphèmes*, Presses Universitaires de Provence.

Vendryès J., 1968, *Le langage. Introduction linguistique à l'histoire*, (1e éd. : 1923), Albin Michel, Paris.

SITOGRAPHIE

www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf

http://www.alfabet.24on.pl/index.php?title=Wyraz

http://encyklopedia.pwn.pl/

www.larousse.fr/dictionnaires

www.oed.com

www.oxforddictionaries.com/definition/english/word

www.wordreference.com/definition/Word

www.macmillandictionary.com/dictionary/british/word_1

www.semantique-gdr.net/dico/index.php/Ambigu%C3%AFt%C3%A9

LE MOT, L'ENTITÉ NOMMÉE ET LES DÉFINITIONS STIPULATIVES

Résumé

Le but de cet article est d'explorer les difficultés que pose la définition du 'mot'. Elles proviennent toutes du fait que les définitions existantes qu'on en donne sont descriptives. En effet, elles prennent en compte une très longue liste de contextes dans un grand nombre de langues où ce terme a été utilisé. Or, aucune des définitions envisagées ne correspond à la totalité de ses usages. Au contraire, l'appelation 'entité nommée', récemment introduite par le biais d'une définition stipulative, est celle qui impose une signification précise dans un contexte donné. Elle échappe donc à l'opposition entre "correcte" ou

"incorrecte". Elle a été créée pour être utilisée dans des situations précises : requête d'informations structurées simples à partir de textes non structurés (ou faiblement formatés), réponse aux questions, résumé de textes, indexation, annotation, etc.

Mots-clés : entité nommée, mot, définition descriptive, définition stipulative

WORD, NAMED ENTITY AND STIPULATIVE DEFINITIONS

Summary

The purpose of this paper is to explore difficulties in defining the term 'word'. They stem from the fact that the existing definitions are descriptive, which means that they take into account a very long list of contexts – linguistic and others – and a variety of languages where this term has been used. None of the definitions considered corresponds to the totality of uses of this term. On the contrary, the term 'named entity', recently introduced by a stipulative definition, is the one in which this new term is given a specific meaning in a given context. It is neither "correct" nor "incorrect". It was created in such a way that it can be useful for its intended purposes such as getting simple structured information out of unstructured (or loosely formatted) text, question answering, classifying information, automatic (or semi-automatic) text summarization, indexing, annotation etc.

Key words: named entity, word, descriptive definition, stipulative definition