

La construction de dictionnaires à partir de l'analyse informatisée de corpus bruts : un outil pour le langagier

Sylvain DELISLE

Université du Québec à Trois-Rivières, Canada

1. Introduction

De nombreuses applications en ingénierie linguistique exigent la construction et la mise à jour de dictionnaires. Bien que les dictionnaires généraux soient largement disponibles, et de plus en plus sur un support adapté au traitement informatique (p. ex. CD-ROM), il en va tout autrement des dictionnaires spécialisés. Ces derniers couvrent habituellement une langue de spécialité, comme celle du monde médical ou de l'informatique, et s'adressent à un public beaucoup plus restreint. Le langagier est appelé à construire des dictionnaires spécifiques à partir d'un corpus donné, soit pour analyser un phénomène linguistique pointu, soit pour caractériser ou modéliser un (sous) langage de spécialité. Le corpus sert alors de source pour construire (ou « dériver ») le dictionnaire.

La construction d'un dictionnaire à partir d'un corpus est une tâche ardue et fastidieuse, surtout lorsque le corpus est de taille ou de complexité importantes. La plupart du temps, le langagier doit effectuer ce travail manuellement ou encore avec des moyens informatiques plutôt rudimentaires. Nous proposons une méthode qui permet de construire facilement un dictionnaire de base à l'aide d'un outil informatisé et ce, directement à partir d'un corpus composé de textes bruts ne nécessitant aucun traitement préliminaire. Cette méthode est indépendante du domaine du texte, est axée sur le verbe et ses arguments, et est basée sur l'analyse syntaxique (automatique) et sémantique (semi-automatique) de corpus bruts. Notre méthode a été implémentée et testée sur des textes anglais de nature technique. Elle est exportable à d'autres langues pour lesquelles des ressources similaires à celles décrites ci-dessous sont également disponibles.

Implémenté surtout en Prolog, le système CASSCAD¹ constitue un environnement informatique qui guide et supporte la tâche de l'utilisateur (c.-à-d. le langagier) pendant la construction d'un dictionnaire ou pendant l'étude de phénomènes linguistiques particuliers. Les composantes du système sont organisées en trois sous-systèmes indépendants mais complémentaires : *i*) un concordancier, *ii*) un sous-système d'analyse lexicale « multi-source » et *iii*) un sous-système d'analyse syntaxique et sémantique.

Le sous-système d'analyse lexicale multi-source intègre plusieurs modules qui ont pour fonction de présenter à l'utilisateur des informations de base sur chacun des mots trouvés dans le corpus (à partir du concordancier) afin d'en construire l'entrée à ajouter au dictionnaire spécifique en cours de construction : liste de concordances et fréquences ; fonctions grammaticales potentielles selon le dictionnaire général *The Collins* ; définition selon la base de données lexicale *WordNet*, à caractère général elle aussi ; et fonction grammaticale la plus probable selon un étiqueteur statistique. L'utilisateur contrôle la construction de chaque nouvelle entrée lexicale à partir de ces diverses sources d'information. Quant au sous-système d'analyse syntaxique et sémantique, il permet de compléter les entrées verbales créées précédemment au cours de l'analyse lexicale. Ces deux analyseurs permettent d'identifier les patrons d'occurrence syntaxiques et les patrons d'occurrence sémantiques (thématiques) pour chaque verbe du corpus. L'analyseur sémantique compile également les fréquences d'occurrence de ces deux types de patrons. De plus, les jeux d'étiquettes thématiques sont adaptables aux exigences du langagier.

Cet article présente, à travers le système CASSCAD, l'essentiel de l'approche que nous proposons pour la construction semi-automatique d'un dictionnaire spécifique. Mais tout d'abord, voyons un exemple d'application en ingénierie linguistique pour lequel, trop souvent, le langagier est dépourvu d'outil informatique adéquat.

2. Exemple d'une approche manuelle en traduction spécialisée

Nous présentons ici un exemple d'application en ingénierie linguistique : la traduction spécialisée. Dans son ouvrage sur la traduction médicale, de la langue anglaise vers la langue française, Rouleau (1994 : 198-199) explique ainsi son analyse du terme 'traitement' :

Rien grammaticalement ne régit l'utilisation d'une préposition particulière avec le verbe « traiter ». Il n'y aurait, selon toute apparence, aucune faute à utiliser « avec », « à » ou « par », si ce n'est que **l'usage a peut-être ses préférences**. Le seul moyen de connaître cet usage, c'est de lire des documents écrits par des spécialistes francophones et d'être attentif aux façons de dire des auteurs. Après avoir dépouillé 13 chapitres écrits par au moins autant de médecins et avoir relevé toutes les phrases où se rencontraient les mots « traitement », « traiter », « thérapie » (corpus de plus de 300 phrases), il est possible d'affirmer que « traiter par » ou « traitement par » est la tournure la plus utilisée.

Puis Rouleau (1994 : 199-203) termine son analyse du terme 'traitement' en présentant une liste de ses cooccurents qui peuvent être soit un nom, p. ex. « traitement

¹ CASSCAD est un acronyme construit à partir d'un réordonnement des lettres soulignées dans « Analyse de Concordance et Analyse Syntaxique et Sémantique pour la Construction de Dictionnaires »

d'urgence », « période de traitement » ; soit un adjectif, p. ex. « traitement anti-tuberculeux », « traitement préventif » ; soit un verbe, p. ex. « le traitement supprime », « prescrire le traitement » ; soit une préposition ou une locution, p. ex. « traitement par [nom d'un médicament ou voie d'administration] ». Fait remarquable, tout ce travail d'analyse de Rouleau est basé *sur des données colligées manuellement, c'est-à-dire, sans aucun outil informatique*. Dans cet article, nous présentons une approche semi-automatique qui vise justement à supporter le travail du langagier en facilitant la collecte de telles données et en permettant de construire un dictionnaire simple à partir du corpus analysé. Cette approche, dérivée de travaux en acquisition automatique de connaissances à partir de textes (Delisle, 1994), possède l'avantage d'offrir beaucoup plus qu'un simple concordancier, comme nous le verrons plus loin.

3. Approche semi-automatique : architecture et fonctions de CASSCAD

L'approche que nous proposons a pour but de rendre plus performante la construction d'un dictionnaire spécialisé ou, encore, spécifique d'un corpus particulier. Dans le cadre du présent article, nous entendons par dictionnaire une base de données fondamentales sur le vocabulaire d'un corpus – la nature exacte de ces données est présentée aux sections 3.2. et 4. Cette base de données est construite à partir du corpus analysé à l'aide du système CASSCAD. Ce système utilise plusieurs logiciels et sources d'information complémentaires. Chacun de ces éléments apporte sa contribution aux trois phases qui constituent l'approche en question.

3.1. Phase 1 – Analyse de concordance (entrée : texte/corpus brut ; sortie : statistiques diverses et liste des concordances)

Le concordancier que nous utilisons est un programme développé à l'UQTR (Boisvert, 1989) auquel nous avons apporté quelques modifications. Écrit en Pascal, ce concordancier est relativement standard et possède, entre autres, les options suivantes :

- fonctionnement en interactif ou par lots ;
- possibilité d'identifier une liste (fichier) de mots à rejeter, c.-à-d. pour lesquels on ne veut pas de concordances ;
- possibilité d'identifier une liste (fichier) de mots exclusifs, c.-à-d. les seuls pour lesquels on veut des concordances. Ces mots peuvent être identifiés à l'aide du symbole '*' (*joker*) : p. ex., ordinateur* couvrira les occurrences de 'ordinateur' et 'ordinateurs'. De plus, il est possible de demander des cooccurrences de N mots. Si un segment de la liste des mots exclusifs compte N mots, le concordancier trouvera *dans une même phrase* les cooccurrences de ces N mots (même si ces N mots ne sont pas contigus) ;
- possibilité de préciser un contexte maximum fixe (de 1 à 9 mots) avant et après chaque occurrence ou, un contexte flottant borné par le début et la fin de la phrase dans laquelle l'occurrence a été trouvée ;
- possibilité de limiter la taille des mots qui seront considérés par le concordancier (p. ex. seulement les mots ayant entre 4 et 20 caractères).

La sortie du concordancier est composée d'abord d'une série de statistiques simples telles que le nombre total de mots, de phrases et de paragraphes du texte ana-

lysé, le nombre moyen de mots par phrase et le nombre moyen de mots par paragraphe, etc. Viennent ensuite les concordances elles-mêmes qui sont ordonnées alphabétiquement. Voici un exemple obtenu à partir d'un texte anglais pour les mots 'change' (2 occurrences), 'changes' (1 occurrence), et 'chapter' (1 occurrence). Le contexte (maximum) est de 5 mots. Chaque entrée est délimitée par « ==> mot » et « > mot [fréquence] ».

```
==> change
later you'll learn how to change some of these defaults
should you can change this to descending order by
> change [2]
==> changes
statements if you have no changes to make to them.
> changes [1]
==> chapter
in this chapter you've learned how to produce
> chapter [1]
```

3.2. Phase 2 – Analyse lexicale multi-source (entrée : liste des concordances ; sortie : dictionnaire de base dico_spé)

Le programme de contrôle de l'analyse lexicale est écrit en C. Il accepte en entrée une liste de concordance sous le format décrit ci-dessus. Pour chaque index (mot) de la liste de concordance, le programme de contrôle présente à l'utilisateur la liste des concordances de ce mot, consulte *The Collins* (Karp *et al.*, 1992) et *WordNet* (Miller, 1990)², et présente à l'utilisateur l'information associée à ce mot, et, finalement, guide l'utilisateur dans la construction d'une entrée de dictionnaire spécifique et adaptée au traitement de la phase 3 – nous qualifierons ce dictionnaire de « spécifique », appelé dico_spé.

L'utilisateur a aussi la possibilité de sauter au mot suivant s'il préfère ne pas traiter un certain mot, par exemple si ce mot appartient à une catégorie grammaticale fermée (d'autant plus que l'analyseur syntaxique de la phase 3 possède son propre dictionnaire, lequel est particulièrement axé sur les catégories grammaticales fermées). Habituellement, les mots des catégories grammaticales ouvertes – nom, verbe, adjectif, adverbe – sont ceux qui présentent le plus grand intérêt pour le langagier. Soulignons également que *WordNet* est utile à cet égard puisqu'il ne porte que sur les mots de la langue anglaise appartenant aux quatre catégories grammaticales ouvertes que nous venons d'identifier.

Nous extrayons du dictionnaire *The Collins* les catégories (ou fonctions) grammaticales potentielles que peut jouer le mot. À l'aide de cette information et des concordances, l'utilisateur peut identifier la (ou les) fonction(s) grammaticale(s) parti-

² Il s'agit là de deux ressources du domaine public. Il existe de nombreuses autres ressources mais pas forcément aussi largement diffusées ou aussi peu coûteuses (p. ex. *The Collins COBUILD English Language Dictionary*, *The Longman Dictionary of Contemporary English*). Voir aussi De Bessé (1991) pour une liste de plusieurs ressources pour la langue française et la langue anglaise.

culière(s) de ce mot dans le corpus sous analyse. Par exemple, le mot anglais 'file' peut fonctionner en tant que nom ou en tant que verbe : c'est ce que *The Collins* nous dit. Cependant, grâce aux concordances, l'utilisateur est à même de constater que dans le corpus sous analyse – on le suppose ici – le mot 'file' n'est utilisé que dans sa fonction de nom³. Ainsi, il sera possible de ne construire que le nombre minimal d'entrées du dictionnaire *dico_spé*, ce qui permet de cerner avec précision le vocabulaire d'un corpus, en plus de contribuer à réduire les problèmes d'ambiguïté lexicale pour le traitement subséquent de la phase 3. Si nécessaire, l'utilisateur pourra même demander au programme de contrôle de soumettre une phrase particulière, tirée d'une occurrence du mot considéré, à l'étiqueteur de Brill (1992) afin de vérifier la catégorie grammaticale probable du mot en question.

La base de données lexicales *WordNet* nous offre une panoplie d'informations. Actuellement, le programme de contrôle de l'analyse lexicale extrait, par défaut, les synonymes et les hyperonymes du mot considéré. Il est cependant possible pour l'utilisateur d'avoir accès aux autres catégories d'information de *WordNet* grâce au programme de contrôle. La mise en parallèle de l'ensemble des synonymes et des concordances d'un mot permet à l'utilisateur de vérifier le sens de ce dernier et, si désiré, de sélectionner la définition⁴ appropriée pour l'inclure dans l'entrée de dictionnaire *dico_spé*. Une fois le sens déterminé, l'accès à l'ensemble des hyperonymes permet d'identifier une catégorie taxonomique pour le mot en question. Poursuivons notre exemple avec le mot anglais 'file' qui est utilisé en tant que nom. L'examen de nos concordances et l'accès aux informations de *WordNet* nous permettent de choisir la définition du sens « data file » et la catégorie taxonomique (hyperonyme) « record ». De plus, on pourra indiquer si le mot est de spécialité ou non. Supposons que tel est le cas, alors l'utilisateur pourra, en interagissant avec le programme de contrôle de l'analyse lexicale, participer à la création de l'entrée *dico_spé* suivante pour le mot 'file' :

```
dico_spe( file, countnoun, sg, _, ['data file'/record/157/special] ).
```

Cet exemple, dans lequel 'countnoun' signifie 'nom nombrable' et 'sg' signifie 'singulier', illustre le format Prolog utilisé par le sous-système d'analyse syntaxique (voir 3.3.). La définition des paramètres de ce format des entrées lexicales est la suivante : 1) mot, 2) catégorie grammaticale, 3) premier paramètre spécifique de la catégorie grammaticale du mot, 4) deuxième paramètre spécifique de la catégorie grammaticale du mot ('_' veut dire sans valeur), 5) liste des définitions, catégories taxonomiques, numéros de concordance⁵ et indicateurs d'appartenance du mot à une langue de spécialité. Mentionnons au passage qu'il n'est pas nécessaire d'entrer toutes les formes d'un mot puisque le sous-système de la phase 3 possède un analyseur morphologique qui permet de reconnaître 'files' comme la forme pluriel du nom 'file', par exemple.

3 Le mot anglais 'file' possède deux fonctions grammaticales (ainsi que plusieurs sens distincts), celles de nom (p. ex. *dossier*) et de verbe (p. ex. *classer*). *CLASSO* permet de distinguer ces cas.

4 Il ne s'agit pas à proprement parler d'une définition exhaustive mais plutôt d'un ou plusieurs mots qui font référence à des concepts clés permettant de distinguer les différents sens d'un mot. L'utilisateur peut accepter ou modifier la définition fournie par *WordNet*, elle sera conservée dans le *dico_spé*.

5 Ce numéro de concordance permet d'associer une occurrence représentative à la définition du *dico_spé*. Ainsi, l'utilisateur peut facilement retracer la provenance de ce mot dans la sortie du concordancier et, par le fait même, dans le corpus.

Si un mot est inconnu, c'est-à-dire s'il n'apparaît ni dans *The Collins* ni dans *WordNet*, il s'agit probablement d'un mot de spécialité pour lequel l'utilisateur devra identifier de façon interactive la valeur des paramètres de son entrée lexicale dans le dictionnaire *dico_spé*. Dans ce cas, le paramètre de spécialité prendrait la valeur de 'special'. Par exemple, dans la phrase « Erythrocyte size and hemoglobinization can be estimated visually on stained films of the blood or can be calculated quantitatively from the hemoglobin, erythrocyte count, and packed cell volume », tiré de Rouleau (1994 : 279), le mot 'hemoglobinization' est inconnu du *The Collins* et de *WordNet* (mais pas 'erythrocyte' !). Dans ce cas, CASCAD demandera à l'utilisateur, par un menu simple, de compléter l'entrée lexicale du mot 'hemoglobinization' – notons aussi que l'étiqueteur peut faire des suggestions utiles à cet effet. Pour ce qui est du traitement de la troisième phase, le cinquième paramètre (définition, catégorie taxonomique, numéro de concordance et indicateur de spécialité) n'est pas obligatoire. L'utilisateur peut donc le laisser indéterminé et le reconsidérer plus tard s'il le désire.

Finalement, si un mot possède plus d'un sens dans le corpus tout en appartenant à une même catégorie grammaticale, CASCAD permet de les distinguer. Par exemple, si le mot 'file' est aussi utilisé dans le sens de meuble (classeur), l'entrée du *dico_spé* sera augmentée comme suit :

```
dico_spe( file, countnoun, sg, _,      ['data file'/record/157 special,
                                       'file cabinet'/'office furniture'/
                                       /294/_ ] ).
```

On remarque que le numéro de concordance nous aidera alors à distinguer les deux sens du mot 'file'. Cette information sera également utile pour compléter les entrées des verbes (section 4.3.). Ici, la distinction des différents sens d'un mot doit être contrôlée par l'utilisateur. D'autres travaux (Yarowsky, 1995 ; Chakravarthy, 1995) se sont intéressés davantage à désambigüiser automatiquement les différents sens d'un mot dans un corpus et ce, à l'aide de ressources semblables à celles que nous utilisons dans notre approche.

3.3. Phase 3 – Analyse syntaxique et sémantique (entrée : texte/corpus brut et dictionnaire *dico_spé* ; sortie : dictionnaire augmenté des entrées de verbes)

L'analyseur syntaxique, nommé DIPETT, et l'analyseur sémantique, nommé HAIKU, sont implémentés en Quintus Prolog 3.2 et en SISCTus Prolog 2.1(#9) sur des stations de travail Sun. La phase 3 commence le traitement du corpus par l'analyse syntaxique du texte original avec l'analyseur syntaxique DIPETT (*Domain-Independent Parser for English Technical Texts*). Pour chaque phrase du corpus, ce parseur produit un arbre d'analyse auquel l'utilisateur pourra apporter des modifications simples à l'aide du module de rattachement (Delisle, 1995), si cela devait s'avérer nécessaire, par exemple, pour corriger l'attachement d'un syntagme prépositionnel⁶.

6 Terry Copeck, un membre du groupe de recherche KAML de l'Université d'Ottawa, participe activement à la réalisation de ce module de rattachement

Vient ensuite l'analyse sémantique semi-automatique effectuée par le module HAIKU. L'arbre d'analyse syntaxique produit par DIPETT est maintenant décomposé par HAIKU et ce dernier détermine les relations sémantiques qui lient ses composants et ce, à trois niveaux complémentaires qui sont associés à autant d'étapes de traitement dans HAIKU⁷. D'abord, les relations entre les propositions qui forment une phrase complexe : par exemple, une proposition peut exprimer une relation de causalité par rapport à une autre proposition de la même phrase (Barker & Szpakowicz, 1995). Ensuite, les relations entre le verbe principal de chaque proposition et ses arguments : il s'agit cette fois d'une analyse au point de vue des Cas sémantiques ; et finalement, les relations entre les éléments des groupes nominaux complexes – ces derniers travaux sont en cours. Dans le présent article, nous insistons davantage sur la deuxième étape, soit l'analyse Casuelle.

3.3.1. Quelques détails sur l'analyseur syntaxique

L'analyse syntaxique nous permet, entre autres, d'accéder aux patrons syntaxiques dont nous avons besoin pour l'analyse sémantique subséquente avec HAIKU. Par opposition à d'autres stratégies de passage plus superficielles, DIPETT effectue une analyse syntaxique détaillée et indépendante du domaine du corpus en entrée : tous les détails sur DIPETT, de même que de nombreuses références pertinentes, apparaissent dans Delisle & Szpakowicz (1991), Copeck *et al.* (1992), Delisle (1994) et Delisle & Szpakowicz (1995). En fait, DIPETT analyse un corpus brut contenu dans un simple fichier texte et dont les mots n'ont pas été lexicalement étiquetés ou annotés au préalable – l'analyse lexicale nous assure que tous les mots du texte à analyser possèdent une entrée dans le *dico_spé* utilisé par DIPETT pour ses informations grammaticales. DIPETT accepte en entrée une chaîne de caractères, une phrase ou un fragment, selon le cas, et produit en sortie un arbre d'analyse unique. Cet arbre unique n'est évidemment pas toujours parfait : le parseur utilise ses heuristiques pour produire son analyse mais, comme il n'a accès à aucune donnée sémantique, il peut construire un arbre plus ou moins correct du point de vue sémantique. C'est pourquoi la fonctionnalité du module de rattachement sera utile à cet égard.

DIPETT tente d'abord de trouver une analyse complète pour chaque phrase soumise en entrée. Lorsque cela est impossible, soit parce que la phrase est grammaticalement incorrecte ou qu'elle est extra-grammaticale par rapport à la grammaire de DIPETT ou, encore, que le temps alloué pour l'analyse d'une phrase est écoulé, l'analyseur tente alors de trouver une analyse en fragments. Il essaie de reconnaître les principales sous-structures de la phrase telles que syntagmes verbaux, syntagmes nominaux, syntagmes prépositionnels, syntagmes adverbiaux ou adjectivaux. Nous considérons qu'il est préférable d'avoir une analyse partielle que rien du tout. D'ailleurs, pour la construction du *dico_spé*, l'analyse fragmentaire permet de répondre à nos objectifs initiaux sans perte importante, car ce sont les structures prédicat-arguments⁸ qui importent, et nous les obtenons par cette analyse par fragments. DIPETT constitue un environnement d'analyse syntaxique qui témoigne de l'im-

7 Les étapes 1 et 3 de HAIKU sont la contribution de Ken Barker du Département d'informatique de l'Université d'Ottawa

8 La pertinence des structures prédicat-arguments pour le traitement informatisé du texte semble avoir effectué un retour en force. Voir à ce sujet Marcus *et al.* (1994) et Grishman (1994)

portance accordée à l'aspect ingénierie du langage dans nos travaux. Des tests avec la version la plus récente du parseur (v3.0) nous donnent les résultats suivants : jusqu'à 95 % des phrases d'un corpus (anglais, technique) sont analysées : 60 % d'analyses complètes et 35 % d'analyses par fragments.

3.3.2. Quelques détails sur l'analyseur sémantique

Nous traitons ici de la partie principale de l'analyse sémantique, c.-à-d. l'analyse Casuelle semi-automatique et interactive – les fondements de cette analyse sont présentés dans Delisle (1994) et *Delisle et al.* (à paraître). Les Cas (Fillmore, 1968 ; Somers, 1987), représentent les relations sémantiques entre le verbe principal d'une proposition et ses arguments syntaxiques, c.-à-d. le sujet, l'objet, les syntagmes prépositionnels et les adverbes. Les relations nommées par les Cas correspondent à des rôles dans l'action associée au verbe. Par exemple, le Cas Agent identifie l'instigateur de l'action. Les Cas se retrouvent dans la syntaxe comme des structures prédicat-arguments dans lesquelles chaque Cas est dénoté par un marqueur et réalisé par un syntagme. Ainsi, dans la phrase « Maxime a réparé sa voiture avec ses nouveaux outils », le Cas Agent est associé à « Maxime », le Cas Objet est associé à « sa voiture » et le Cas Instrument est associé à « ses nouveaux outils ». Nous avons opté pour une analyse sémantique basée sur les Cas pour deux raisons majeures. Premièrement, l'analyse Casuelle permet d'établir un lien explicite entre la syntaxe et la sémantique ; ceci est essentiel dans une approche basée sur la syntaxe. Deuxièmement, l'analyse Casuelle s'effectue en des termes relativement simples et intuitifs qui en rendent les concepts accessibles à l'utilisateur ; il s'agit là d'un point important dans le contexte d'une approche semi-automatique orientée vers le langage.

Nous avons construit un système de Cas général et indépendant de tout domaine particulier. C'est ce système de Cas qui, par défaut, est utilisé par HAIKU. Il comporte 28 Cas regroupés en 5 catégories (les abréviations des Cas apparaissent après le '/') : 1) PARTICIPANT : Agent/agt, Beneficiary/benf, Experiencer/expr, Instrument/inst, Object/obj, Recipient/recp ; 2) CAUSALITY : Cause/caus, Effect/eff, Opposition/opp, Purpose/purp ; 3) TIME : Frequency/freq, Time_at/tat, Time_from/tfrm, Time_to/tto, Time_through/ttru ; 4) SPACE : Direction/dir, Location_at/lat, Location_from/lfrm, Location_to/lto, Location_through/ltru, Orientation/ornt ; 5) QUALITY : Accompaniment/acmp, Content/cont, Exclusion/excl, Manner/man, Material/matr, Measure/meas, Order/ord. La justification et la définition de ces Cas sont présentées dans Barker et al. (1993). Ce sont les abréviations de ces Cas qui serviront à la construction des patrons sémantiques.

4. La construction des entrées complémentaires pour les verbes

L'analyseur Casuel de HAIKU accepte comme entrée un arbre d'analyse produit par DIPETT et y associe, semi-automatiquement, les patrons Casuels qui représentent le mieux le sens de la phrase⁹. Les Cas sont réalisés dans la syntaxe de deux façons : 1) de façon lexicale (c.-à-d. par un marqueur explicite dans la syntaxe de surface), par

⁹ Lorsque la phrase contient plusieurs propositions, HAIKU la découpe en une suite de propositions qui sont analysées les unes à la suite des autres

exemple lorsqu'une préposition introduit un syntagme prépositionnel, et 2) de façon *positionnelle*, par un marqueur implicite associé au sujet (psubj), à l'objet direct (pobj) ou à l'objet indirect (piobj). Tout comme pour les Cas ci-dessus, ce sont les symboles associés à ces marqueurs qui serviront à la construction des patrons. Par exemple, le patron syntaxique (PSY) psubj-pobj-at est associé à une proposition dans laquelle le verbe principal possède un sujet, un objet direct et un syntagme prépositionnel introduit par la préposition 'at'. De même, le patron sémantique (PSÉ) agt-obj-lto peut être associé à une proposition dont le PSY est psubj-pobj-at et dans laquelle le verbe principal possède un sujet qui tient le rôle agent, un objet direct qui tient le rôle objet et un syntagme prépositionnel qui tient le rôle de location_{to} (destination)¹⁰.

L'analyseur Casuel effectue un type d'apprentissage automatisé qui possède les trois principales caractéristiques de l'apprentissage basé sur les occurrences (ou *instance-based learning*, voir Aha *et al.*, 1991) : *i*) c'est un apprentissage supervisé, c.-à-d. contrôlé par l'utilisateur ; *ii*) c'est un apprentissage incrémentiel ; et *iii*) c'est également un apprentissage basé sur les similarités entre le nouveau patron à identifier et ceux déjà assimilés. Pour ce faire, HAIKU utilise quatre dictionnaires simples (voir 4.1. à 4.4.) qui peuvent être vides au début de l'analyse d'un corpus. Dans ces circonstances, la contribution de l'utilisateur sera plus importante initialement et s'allègera à mesure que HAIKU garnira ses dictionnaires de façon incrémentielle. Ce sont ces quatre dictionnaires construits par HAIKU qui viendront compléter le dico_{spé} initial résultant des phases 1 et 2.

Tous les dictionnaires de HAIKU sont continuellement mis à jour pendant l'analyse du corpus. Chaque proposition se voit attribuer un PSÉ unique à la suite de l'intervention de l'utilisateur, soit qu'il approuve la suggestion du système, soit qu'il la modifie. Pour faire une suggestion à l'utilisateur, l'analyseur Casuel fouille ses dictionnaires dans le but de trouver un PSÉ qui correspond le mieux au PSY de la proposition considérée. Pour ce faire, on utilise un algorithme de contrôle de l'analyse Casuelle (Delisle *et al.*, à paraître) couplé à un algorithme simple de filtrage (Delisle *et al.*, 1993) qui permet de trouver le ou les meilleurs PSÉ candidats en fonction du PSY – l'analyseur Casuel utilise aussi la phrase exemple conservée dans le dictionnaire cmpDict afin d'illustrer la situation à l'utilisateur et ainsi simplifier sa décision.

Si le ou les PSÉ suggéré(s) par le HAIKU ne semblent pas acceptables à l'utilisateur, ce dernier est alors appelé à intervenir en identifiant lui-même le PSÉ approprié. HAIKU affiche des informations complémentaires comme la liste des Cas associés aux marqueurs de Cas de la proposition analysée et la liste des Cas manipulés par le système. Notons que l'utilisateur peut ajouter de façon dynamique ses propres Cas à l'ensemble des 28 Cas prédéfinis dans le système à tout moment au cours de son interaction avec HAIKU. Ce dernier permet également de sauver automatiquement la liste de Cas de l'utilisateur afin d'en faciliter la réutilisation.

Voyons maintenant la structure de ces quatre dictionnaires. Pour illustrer nos propos, nous utiliserons le petit corpus suivant à titre d'exemple : « Bob printed the

¹⁰ L'ordre n'importe pas dans les patrons, seule leur interprétation sémantique est importante. Ainsi, psubj-pobj-at-by est équivalent à psubj-pobj-by-at, et, de façon similaire agt-obj-lat-tat est équivalent à agt-obj-tat-lat.

new data file. Beth and Tom will print their letters. Their boss could not print the production report on the new laser printer. The new computer caused a power failure yesterday. We know that your boss would not delete all your data. These new employees have deleted my letters from my disk. »

4.1. mDict (*meaning dictionary*)

Le dictionnaire de sens (mDict) contient des entrées pour les mots individuels : verbes, prépositions ou adverbes. Pour les deux dernières catégories, le mDict contient la liste fixe des Cas qui peuvent être marqués par ces mots (voir Barker *et al.*, 1993). Pour les verbes, une entrée contient : 1) la liste des PSY trouvés dans le corpus, ainsi que le nombre d'occurrences de chaque PSY ; et 2) la liste des Cas qui ont été associés à chaque marqueur de Cas, ainsi que le nombre d'occurrences de chaque association. Voici le contenu intégral en Prolog des entrées des verbes du mDict après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le mDict était vide au départ) :

```
mDict(cause, ['psubj-pobj-adv':1],
        [[adv, [tat:1]], [pobj, [obj:1]], [psubj, [agt:1]]]).
mDict(delete, ['psubj-pobj':1, 'psubj-pobj-from':1],
        [[from, [lfrm:1]], [pobj, [obj:2]], [psubj, [agt:2]]]).
mDict(know, ['psubj-pobj':1],
        [[pobj, [obj:1]], [psubj, [agt:1]]]).
mDict(print, ['psubj-pobj':2, 'psubj-pobj-on':1],
        [[on, [lto:1]], [pobj, [obj:3]], [psubj, [agt:3]]]).
```

4.2. cmpDict (*Case-marker pattern dictionary*)

Le dictionnaire des patrons de marqueurs de Cas (cmpDict) contient une entrée pour chaque PSY. Chaque entrée associe au PSY la liste des PSÉ qui ont été attribués à ce PSY au cours de l'analyse du corpus, de même que le nombre d'occurrences de chacun des PSÉ. De plus, chaque PSÉ est illustré par une phrase exemple, tirée du corpus analysé, que l'utilisateur aura considérée comme représentative du PSÉ en question. Voici le contenu intégral en Prolog des entrées des verbes du cmpDict après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le cmpDict était vide au départ) :

```
cmpDict('psubj-pobj', [['agt-obj':4,
        '''[bob,printed,the,new,data,file,.]''' ]]).
cmpDict('psubj-pobj-adv', [['agt-obj-tat':1,
        '''[the,new,computer,caused,a,power,failure,
        yesterday,.]''' ]]).
cmpDict('psubj-pobj-from', [['agt-obj-lfrm':1,
        '''[these,new,employees,have,deleted,my,letters,from,
        my,disk,.]''' ]]).
cmpDict('psubj-pobj-on', [['agt-obj-lto':1,
        '''[their,boss,could,not,print,the,production,
        report,on,the,new,laser,printer,.]''' ]]).
```

4.3. cpDict (*Case pattern dictionary*)

Le dictionnaire des PSÉ (cpDict) contient une entrée pour chaque PSÉ rencontré dans le corpus et lui associe la liste des verbes qui se sont vus attribuer un tel PSÉ. Voici le contenu intégral en Prolog des entrées des verbes du cpDict après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le cpDict était vide au départ) :

```
cpDict('agt-obj', [delete, know, print]).
cpDict('agt-obj-lfrm', [delete]).
cpDict('agt-obj-lto', [print]).
cpDict('agt-obj-tat', [cause]).
```

Notons qu'il est possible de distinguer plus finement les verbes en associant à ceux-ci le numéro de concordance introduit à la section 3.2. Par exemple, supposons que les deux PSÉ associés au verbe 'delete' correspondent à autant de sens et que nous désirions les démarquer. Les deux premières entrées du cpDict pourraient alors être comme suit :

```
cpDict('agt-obj', [delete/489, know, print]).
cpDict('agt-obj-lfrm', [delete/502]).
```

4.4. ccvpIndex

Le dernier dictionnaire sert en fait de structure indexée afin de faciliter l'accès aux résultats produits par la phase 3 et conservés dans un fichier de sortie qui est indépendant des quatre dictionnaires dont il est question ici. Dans ce fichier de sortie, on retrouve deux structures pour chaque unité¹¹ (units dans le ccvpIndex) du corpus analysée par le système : l'arbre d'analyse syntaxique et la structure de Cas de HAIKU. Ces deux structures sont co-indexées grâce à un simple numéro d'identification unique (# dans le ccvpIndex). Le ccvpIndex met donc en association toutes les occurrences distinctes de PSY, PSÉ, verbe, et autres détails sur le contexte d'occurrence de ces patrons, c'est-à-dire le numéro de l'unité dans laquelle un PSY, un PSÉ et un verbe particuliers ont été rencontrés dans le corpus ; la sous-catégorisation de surface¹² du verbe (sr_types dans le ccvpIndex) telle que trouvée dans le corpus ; et le temps du verbe dans cette occurrence. Voici le contenu intégral en Prolog des entrées des verbes du ccvpIndex après l'analyse syntaxique et l'analyse Casuelle du petit corpus ci-dessus (le ccvpIndex était vide au départ) :

```
ccvpIndex('psubj-pobj', 'agt-obj', delete,
          units([[#(5), sr_types('np-np'),
                  tense([would_conditional_present_simple]]))]).
ccvpIndex('psubj-pobj', 'agt-obj', know,
```

¹¹ Nous appelons unité chaque segment considéré pour analyse. Un segment peut correspondre à une phrase complète ou à un syntagme isolé (ou une phrase incomplète).

¹² 'np' (noun phrase) veut dire syntagme nominal ; 'nom_cl' (nominal clause) veut dire proposition nominale ; 'adv' signifie groupe adverbial ; et 'on', 'from' désignent un syntagme prépositionnel introduit, respectivement, par la préposition 'on' ou la préposition 'from'.

```

units([[#(5), sr_types('np-nom-cl'),
      tense([infinitive])]]).
ccvpIndex('psubj-pobj', 'agt-obj', print,
  units([[#(1), sr_types('np-np'),
        tense([past_simple]),
        [#(2), sr_types('np-np'),
          tense([future_simple])]]])).
ccvpIndex('psubj-pobj-adv', 'agt-obj-tat', cause,
  units([[#(4), sr_types('np-np-adv'),
        tense([past_simple])]]])).
ccvpIndex('psubj-pobj-from', 'agt-obj-lfrm', delete,
  units([[#(6), sr_types('np-np-from'),
        tense([present_perfect_simple])]]])).
ccvpIndex('psubj-pobj-on', 'agt-obj-lto', print,
  units([[#(3), sr_types('np-np-on'),
        tense([could_conditional_present_simple])]]])).

```

5. Aperçu de quelques travaux connexes

L'extraction automatique ou semi-automatique de connaissances, d'informations ou de données, à partir de textes de tous genres est, depuis la fin des années 80, un domaine de recherche très actif en informatique linguistique. Citons, à titre d'exemple, les travaux récents d'Agarwal (1994), d'Appelt *et al.* (1993), de Gomez *et al.* (1994), d'Ogonowski *et al.* (1994) et de Delisle (1994).

Parmi ces travaux, plusieurs portent sur la construction automatique de dictionnaires et de lexiques avec des objectifs similaires à ceux de l'approche que nous avons décrite dans le présent article. Mentionnons, entre autres, Cardie (1993), qui propose une approche permettant d'acquérir à partir d'un corpus les fonctions grammaticales et les sens des mots appartenant à une catégorie ouverte ; Grishman *et al.* (1994a) et Sanfilippo (1994), qui décrivent certains aspects de la problématique de la construction d'un grand lexique pour des fins de traitement informatique de textes, ainsi que certaines solutions qu'ils proposent ; Riloff (1993) et Soderland *et al.* (1995), qui présentent chacun un système conçu pour la construction automatique d'un dictionnaire spécifique d'un domaine donné, et ce, dans le contexte d'une application en extraction d'information ; et Sanfilippo & Poznanski (1992), qui suggèrent une approche au problème de la mise en correspondance des différents sens d'un mot lorsque ceux-ci proviennent de différents dictionnaires informatisés.

Il existe aussi de nombreux travaux portant davantage, quoique non exclusivement, sur les entrées des verbes de ces dictionnaires. Soulignons, entre autres, Framis (1994), Grishman & Sterling (1994) et Manning (1993), qui présentent des approches à l'identification automatique des restrictions (ou contraintes) de sélection à partir de l'analyse d'un corpus ; Myaeng *et al.* (1994) et Pugeault *et al.* (1994), qui s'intéressent particulièrement à l'extraction des structures prédicat-arguments à partir des textes ; et Basili *et al.* (1992) et Sekine *et al.* (1992), qui proposent des méthodes pour acquérir automatiquement à partir d'un corpus des collocations de nature sémantique.

6. Conclusion

Le contenu des dictionnaires construits par le système CASSCAD peut grandement aider le langagier (p. ex. terminologue ou traducteur) ou l'ingénieur de la connaissance dans la construction d'un dictionnaire spécialisé (ou spécifique) et, de façon plus particulière, dans l'étude des verbes d'un corpus pour en préciser les propriétés syntaxiques et sémantiques. Ainsi, le *dico_spe* nous dit :

- quels mots apparaissent dans un corpus et quelles informations s'y rattachent (catégorie grammaticale, définition, catégorie taxonomique, indicateur de spécialité, etc.) ;

et les quatre dictionnaires construits par HAIKU nous disent, en plus, pour les verbes :

- quels sont leurs patrons syntaxiques et leurs fréquences d'occurrence respectives ;
- quels sont leurs sous-catégorisations de surface ;
- quels sont leurs patrons sémantiques et leurs fréquences d'occurrence respectives ;
- quels verbes ont des patrons (syntaxiques ou sémantiques) identiques ou similaires ;
- dans quelles phrases du corpus apparaissent un verbe, un PSY ou un PSÉ particuliers ?
- dans quelles phrases du corpus apparaît le verbe V avec le PSY ou le PSÉ P ?
- dans quelles phrases du corpus apparaissent ensemble le PSY P1 et le PSÉ P2 ?

Il nous semble qu'un système comme CASSCAD pourrait être d'un grand secours pour le langagier qui souhaite construire une classification de verbes (voir Dixon, 1991 ; Levin, 1993). Dans le futur, nous prévoyons améliorer le traitement des mots composés afin de permettre à l'utilisateur de les considérer comme des unités linguistiques lorsque désiré : les éléments sont en place dans le concordancier, l'analyseur lexical et l'analyseur syntaxique, mais il nous reste à les intégrer de façon cohérente. De plus, il serait avantageux de pouvoir utiliser les catégories taxonomiques tirées de *WordNet* (ou créées par l'utilisateur) afin de rendre plus spécifiques les collocations sémantiques et les contraintes de sélection des verbes accumulées par HAIKU. Nous planifions également une expérimentation sur de gros corpus afin d'évaluer notre approche sur une plus grande échelle : cela permettrait de répondre à des questions comme « quel est la proportion des verbes identifiés lors de l'analyse lexicale qui se voient associer des PSY ou PSÉ une fois la phase 3 complétée ? ». Une autre question intéressante est celle de la généralisation de notre approche : est-elle utile à la construction d'un dictionnaire à caractère général ?

Remerciements

Je remercie tous ceux qui ont contribué aux travaux mentionnés dans cet article : d'abord, René Boisvert pour avoir réalisé le programme de concordance ; ensuite, Georges Diop Rogandji pour avoir implémenté le module d'analyse lexicale multi-source ; puis, tous les membres du groupe de recherche KAML du Département d'informatique de l'Université d'Ottawa qui, au fil des années, ont testé DIPETT et HAIKU sans pitié aucune, en plus d'apporter des idées qui ont contribué à mes recherches. Je remercie aussi le CRSNG (Conseil de Recherches en Sciences Naturelles et Génie du Canada) de son support financier. Finalement, je remercie Maurice Rouleau pour avoir relu cet article.

Réseau notionnel, intelligence artificielle et équivalence en terminologie multilingue : essai de modélisation

Marc VAN CAMPENHOUDT

Centre de recherche TERMISTI, Institut supérieur de traducteurs et interprètes, Bruxelles, Belgique

1. Introduction

L'approche notionnelle constitue l'un des fondements de la terminologie. Depuis plusieurs années, des équipes de recherche ont développé des logiciels permettant de naviguer au travers des réseaux notionnels et ainsi de mieux appréhender la notion au sein de son microdomaine. Des gestionnaires comme *MCA* (Université de Clermont-Ferrand), *Termisti* (ISTI, Bruxelles) ou *Code* (Université d'Ottawa) constituent autant de pas successifs vers la construction de bases de connaissances et vers l'intelligence artificielle.

Cet article a pour principal objectif de montrer que l'exploitation logique des réseaux notionnels au sein de bases de connaissances terminologiques (B.C.T.) multilingues devrait aussi permettre de gérer divers problèmes d'équivalence. Il se fonde sur un corpus d'exemples extraits de *De la quille à la pomme de mâât* (Paasch, 1901), un vaste dictionnaire nautique trilingue dont l'organisation notionnelle exemplaire a conduit à l'ébauche du modèle théorique ici exposé¹. De par sa tâche d'expert maritime, son auteur, le capitaine Heinrich Paasch, a été inévitablement confronté au non-isomorphisme² des langues. Il est très aisé d'affirmer que tel ou tel dictionnaire est fondé sur une approche notionnelle. Rares sont pourtant, à nos yeux, les auteurs de terminographies multilingues qui vont jusqu'au bout de cette logique et distinguent réellement chacune des notions propres à chacune des langues envisagées.

1. La relation hyponymique retiendra plus particulièrement notre attention. La place des autres relations notionnelles dans le modèle a déjà été décrite dans la thèse que nous avons consacrée à ce dictionnaire (Van Campenhoudt, 1994) et dont cet article est issu.

2. À la suite de Lyons (1970 : 45), nous parlerons de (*non-*)isomorphisme entre les langues et de *chevauchement culturel*.

2. Découpage notionnel et confrontation des langues

2.1. La tradition viennoise face à l'équivalence

La linguistique a depuis longtemps montré que toutes les langues n'approchent pas la réalité de la même manière et que de nombreux problèmes se posent lors de l'établissement d'équivalences. Eugen Wüster, le chef de file de l'école viennoise, avait assurément pris conscience du fait que les systèmes de notions varient d'une langue à l'autre. En divers passages de son œuvre³, il a rappelé cet état de fait et regretté que de nombreux terminographes réalisent des œuvres dans lesquelles le système notionnel est conditionné par une langue particulière, ce qui débouche inévitablement sur des impossibilités de traduction.

Face aux problèmes d'équivalence soulevés par la divergence notionnelle entre les langues, Wüster (1971 : 44-45) proposait pour solution d'adopter un système notionnel commun, normalisé⁴ au niveau international. Son principal héritier, Helmut Felber (1987 : 131) ne semble pas échapper à la confusion qui ferait de la terminologie une discipline foncièrement normative, habilitée à déterminer une fois pour toutes ce qui existe et ce qui n'existe pas, soumettant toutes les langues de l'humanité au *diktat* conceptuel de quelques langues européennes. Aujourd'hui encore, Felber (1994 : 165) propose de procéder à une unification notionnelle en cas de non-isomorphisme. Il est pourtant paradoxal que dans le même temps il présente comme normal le fait qu'un même objet puisse être conceptualisé de manière différente selon les disciplines envisagées⁵.

2.2. Un réseau notionnel interlinguistique (R.N.I.)

Dans un article intitulé *Terminological Equivalence and Translation*, Reiner Arntz (1993 : 6-7) se fonde sur le problème de la divergence dans la manière dont les langues désignent les couleurs pour montrer qu'il convient avant toute chose de décrire les systèmes notionnels propres à chaque langue. Pour Arntz, l'approche descriptive constitue le fondement d'une terminologie multilingue orientée vers la traduction. Elle permet de comparer les systèmes notionnels de chaque langue pour découvrir toutes les divergences à prendre en compte lors de l'établissement des équivalences. Il préfère toutefois ne pas recourir à la normalisation dans une perspective de traduction et propose de résoudre les difficultés éventuelles par des procédés linguistiques tels l'emprunt, la néologie et la paraphrase.

Cette perspective est intéressante, car elle consiste à rendre compatibles les réseaux notionnels de chaque langue plutôt que de les standardiser internationalement. Dans une terminographie multilingue, chaque langue doit pouvoir servir indistinctement

3 Lire notamment Wüster (1971 : 36ssq et 44-45, 1968 : 219, 1981 : 66 et 71). Ce constat est également présent chez Felber (1987 : 128ssq).

4 Dans le même article, Wüster (1971 : 40-41) va même jusqu'à parler d'*épuraton*, mot sans ambiguïté quant à la nature de la tâche de normalisation.

5 Felber (1994 : 169) propose une intéressante modélisation de cette variation notionnelle en fonction des disciplines. Il est intéressant de noter que l'auteur ne tient pas compte du cas où la différence de conceptualisation est marquée par un terme différent. Il est vrai qu'un tel cas s'apparenterait étrangement à celui d'une inacceptable différence de découpage notionnel entre les langues.

tement de langue source ou de langue cible. La seule manière de satisfaire à cette exigence sans verser dans la normalisation semble bien être de fusionner les réseaux notionnels de chacune des langues considérées de manière à rendre compte de toutes leurs particularités. Pour établir ce réseau notionnel commun, que nous nommerons dorénavant **réseau notionnel interlinguistique** ou **R.N.I.**, le terminologue doit nécessairement partir de l'observation des désignations de chaque langue pour identifier les concepts qu'elle véhicule (sémasiologie). La recherche des équivalents (onomasiologie) s'effectue ensuite, mais elle doit, autant que possible, être respectueuse des faits décrits.

Dans une telle perspective, l'activité de normalisation n'est pas une condition nécessaire à l'établissement de l'équivalence. Arntz (*ibid.*) décrit d'ailleurs la normalisation terminologique comme une activité parallèle, quand bien même elle est également précédée d'une phase descriptive. Contrairement à ce qu'affirme Felber (1987 : 152), l'approche descriptive n'est donc pas qu'« une phase préliminaire qui prépare le travail terminologique normatif » ; elle peut aussi constituer le fondement d'une démarche d'établissement de l'équivalence.

S'il est arrivé à Wüster (1981 : 79) de parler de « *système de notions international* », il ne semble pas avoir voulu désigner par ces mots la démarche du R.N.I. décrite ci-dessus, mais plutôt les systèmes notionnels unifiés internationalement qui existent comme tels dans quelques domaines et qui ne requièrent donc pas de normalisation. Toutefois, l'introduction du *Dictionnaire multilingue de la machine-outil* montre que, confronté à la réalité des langues, Wüster (1968 : 2.19) a adopté une démarche plus descriptive que normative⁶.

2.3. Principe d'équivalence notionnelle et découpage conceptuel

Dans un article fort intéressant, Bernard Levrat et Gérard Sabah (1990 : 93) rappellent que dans divers réseaux sémantiques, un lien d'équivalence permet de représenter les relations de synonymie. Ils montrent que « lors de la gestion automatique du réseau, ce lien peut être utile pour mettre en évidence des polysémies potentielles : si A est synonyme de B et si A est synonyme de C alors que B n'est pas synonyme de C, c'est que probablement A possède deux sens qui devraient être différenciés par deux nœuds du réseau. »

Les réseaux, qu'ils soient notionnels ou sémantiques, sont bâtis sur une perspective conceptuelle. Cette citation montre que dans un réseau sémantique, la synonymie est basée sur l'équivalence entre deux concepts, comme l'est la traduction dans un réseau notionnel interlinguistique. Tout semble donc nous autoriser à transposer la loi qui vient d'être énoncée pour l'adapter à la distinction des notions (ou concepts) en terminologie traductionnelle. L'énoncé qui suit explique comment identifier les termes qui renvoient à plusieurs notions et qui devront vraisemblablement faire l'objet d'un dégroupement hyponymique au sein du R.N.I. :

6 À ce sujet, lire également Arntz (1993 : 6-7)

Si A de L_1 est équivalent à α de L_2 et si A de L_1 est équivalent à β de L_2 alors que α de L_2 n'est pas synonyme de β de L_2 , c'est que probablement A de L_1 possède deux sens qui devraient être différenciés par deux nœuds du réseau.

	L_1	L_2
notion 1	A	= α
notion 2 :	A	= β

Ce principe, que nous dénommerons **principe d'équivalence notionnelle (P.E.N.)**, est scrupuleusement respecté dans notre corpus de référence. Grâce au dégroupement homonymique, le terminographe a veillé à ce qu'à chaque notion identifiée corresponde un terme adéquat. Ces dégroupements homonymiques peuvent être dus à une ou plusieurs langues :

Watch. The act of vigilance.

Veille Action de veiller.

Wache ; Wachen.

Watch. The divisions of time by day and night on board a ship, when a certain portion of a vessel's crew are on duty

Quart. Division du temps tant le jour que la nuit à bord d'un navire, pendant laquelle une certaine partie de l'équipage est de service sur le pont

Wache. Die Zeiteintheilung bei Tag und Nacht an Bord eines Schiffes, an der ein gewisser Theil der Bemannung Dienst auf Deck hat.

Watch. The men employed to form a watch : for instance . the half of the crew.

Bordée. Nom donné à la partie d'un équipage formant le quart.

Wache. Benennung für die Leute, welche eine Wache bilden (zu einer Wache gehören)

(Paasch 1901 : 576)

Pilotage. The skill or knowledge of a pilot respecting coasts, rivers, channels, currents, etc.

Pilotage. La connaissance d'un pilote des côtes fleuves, courants, etc.

Lootsenkunde. Die Kenntniss eines Lootsen in Betreff der Küsten, Flüsse, Strömungen, des Fahrwassers u s w

[.]

[.]

[]

Pilotage. The money paid for the services of a pilot.

Droits de pilotage. Contributions perçues pour les services rendus par les pilotes

Lootsengeld. Das, für die Dienste eines Lootsen gezahlte Geld.

[...]

[.]

[.]

Pilot-office. The building or the rooms in a sea-port, in which the Pilot-master and assistants conduct the business in connection with pilotage.

Pilotage. Bureaux de l'Administration du Pilotage dans un port, où l'inspecteur du pilotage et ses assistants dirigent les affaires se rapportant au pilotage des navires.

Lootsenwesen. Gebäude, in welchem sich die Büreaus einer Lootsenbehörde befinden und woselbst alle dieses Fach betreffenden Angelegenheiten erledigt werden.

(Paasch 1901 512)

Bien entendu, l'application stricte du principe d'équivalence notionnelle implique que la présence d'un synonyme dans l'une des langues concernées suffise à justifier le principe de dégroupement, conformément à la loi d'établissement des nœuds du réseau monolingue (Levrat et Sabah *op.cit.*). Par exemple, dans le passage suivant :

Breakwater. A structure of timber; iron or steel plates, say from one to four feet in height according to the size of the vessel, fitted across fore-castle-decks (notably of large steamers) to break the force of any sea shipped over the bows.

Brise-lame. Construction en bois, en fer ou en acier, ayant une hauteur de un à quatre pieds selon la grandeur du bâtiment, fixée en travers d'un pont de gaillard (notamment sur les grands steamers) pour briser les lames ou pour diminuer la force de celles-ci lorsqu'elles s'élèvent sur l'avant du navire.

Brechwasser. Ein Gefüge von Planken, eisernen oder stählernen Platten, je nach der Grosse des Schiffes, ein bis vier Fuss hoch, welches quer über em Backdeck (besonders bei grossen Dampfern) angebracht ist, um die Gewalt der über den Bug stürzenden Wellen zu brechen

(Paasch 1901 : 43)

Breakwater. A stone-wall built up from the bottom of the sea, at the entrance of a bight, etc., to form a harbour, or to shelter one

Brise-lames. Sorte de digue (ou mur de pierres) érigée sur le fond de la mer en avant d'un port et qui s'élève jusqu'au-dessus des eaux, pour amortir la violence des vagues, et protéger le port

Wellenbrecher; Brechwasser. Eine am Eingange einer Bucht u.s.w vom Grunde der See aufgebaute, deichähnliche Mauer, an welcher sich die Gewalt der Wellen bricht

(Paasch 1901 : 424)

De nombreux cas de dégroupements homonymiques, tel le dernier cité, apparaissent d'autant plus justifiés que les notions concernées relèvent de sous-domaines différents et ne sont donc pas liées. Très souvent d'ailleurs, la prise en compte des liens notionnels corrobore la nécessité de distinguer plusieurs notions en vertu du P.E.N. Ainsi, il suffit de s'apercevoir que le terme peut être classé dans deux arborescences espèce-genre différentes pour se rendre compte qu'il recouvre vraisemblablement deux notions différentes.

On notera toutefois que des notions se distinguent parfois sur la base du seul réseau notionnel, sans qu'intervienne le P.E.N. : elles sont désignées par des termes homonymes dans chaque langue, mais recouvrent des réalités distinctes, liées par une relation fonctionnelle⁷ qui ne s'exprime pas aisément.

Course. The direction, over sea, from one point of land to another.

Route. Chemin à parcourir par voie de mer, de l'un point de terre à un autre

Kurs; Curs. Die Richtung über See, von einer Landspitze zu einer anderen

Course. The direction in which a vessel sails by compass.

Route. La direction qu'un navire suit d'après la boussole

Kurs; Curs. Der Kompassstrich, auf dem ein Schiff segelt, um einen bestimmten Ort zu erreichen.

(Paasch 1901 : 443)

7. Sur les relations fonctionnelles, lire Levrat (1990).

2.4. Vers une multiplication du nombre de notions ?

Une terminographie multilingue conçue sur la base du réseau notionnel d'une seule langue ne fonctionne correctement que lorsque ladite langue sert de langue source. Le rôle du principe d'équivalence notionnelle est précisément de répondre à une des exigences fondamentales du R.N.I. : que chaque langue puisse indifféremment servir de langue source ou de langue cible. Ce principe a toutefois pour corollaire inévitable un net accroissement du taux d'homonymie pour les langues qui possèdent les notions de plus grande extension. Tel est le cas chez Paasch, puisque comme l'attestent les extraits déjà cités, de nombreux termes homonymes sont présents dans *De la quille à la pomme de mât*.

À travers l'étude de ce dictionnaire, nous avons tenté d'isoler les principes théoriques qui expliquent comment et pourquoi le recours à l'homonymie permet, autant que l'emprunt, la néologie ou la périphrase, de résoudre des problèmes d'équivalence partielle. Nous avons ainsi été amené à découvrir que contrairement à nos prévisions, l'inévitable accroissement du nombre de notions au sein du R.N.I. était souvent restreint par un étrange mécanisme régulateur dont nous nous proposons d'analyser le fonctionnement.

3. Relation d'hyponymie, homonymie et équivalence

L'idée que les relations qui lient les notions d'un même domaine ou sous-domaine forment un réseau porteur d'informations est fort proche de celle qui a conduit à l'élaboration des réseaux sémantiques. La comparaison peut aller beaucoup plus loin, puisque les relations qui entrent en jeu dans les réseaux notionnels et dans les réseaux sémantiques sont de nature voisine. Les cognitivistes, qui ont joué un rôle fondamental dans l'établissement des premiers réseaux sémantiques, ont mis en valeur le rôle fondamental de la relation hyponymique espèce-genre (ci-après, relation TY). À la suite des travaux de Quillian (1967), on pensait copier ainsi un processus cérébral de stockage lexical fondé sur le principe d'héritage des propriétés au sein d'arborescences fondées sur la relation hyponymique⁸. Or, ce type de relation occupe une place prépondérante dans la macrostructure du dictionnaire de Paasch et c'est l'exploitation de quelques principes liés à l'hyponymie qui permet d'y résoudre divers problèmes d'équivalences.

3.1. Trois réseaux notionnels à confronter

Pour montrer la manière dont fonctionne le R.N.I., nous allons isoler, à titre d'exemple, une petite partie du réseau notionnel du sous-domaine de la voileure (Paasch, 1901 : 338-352). Ce domaine se révèle particulièrement intéressant dans la mesure où, pour dénommer des réalités identiques, l'anglais, le français et l'allemand ont adopté des systèmes de désignation fort proches et fondés sur l'hyponymie. Toutefois, diverses divergences de point de vue posent des problèmes d'isomorphisme entre ces trois langues.

⁸ Les terminoticiens s'intéressent beaucoup aux travaux des cognitivistes, notamment à ceux qui ont abouti à la création de *Wordnet* (Miller, 1990).

En anglais, comme en français et en allemand, le système de désignation est fortement motivé, puisque les voiles sont nommées en fonction de leur emplacement. Le tableau n° 1 montre ainsi qu'en français, les voiles carrées se nomment de bas en haut *basses voiles*, *huniers*, *perroquets* et *cacatois*. On distingue le mât sur lequel elles se situent en joignant à leur nom (ci-après *N*) les adjectifs *petit N* (situé sur le mât de misaine), *grand N* (sur le grand mât), *grand N avant* (sur le grand mât avant), *grand N central* (sur le grand mât central) ou *grand N arrière* (sur le grand mât arrière). Pour le mât d'artimon, les désignations sont particulières (de bas en haut : *perroquet de fougue*, *perruche*, *cacatois de perruche* et *contre-cacatois de perruche*). À l'époque considérée, les huniers et les perroquets se subdivisent le plus souvent en deux voiles superposées ; celle du dessous est dite *fixe* et celle du dessus est dite *volante*.

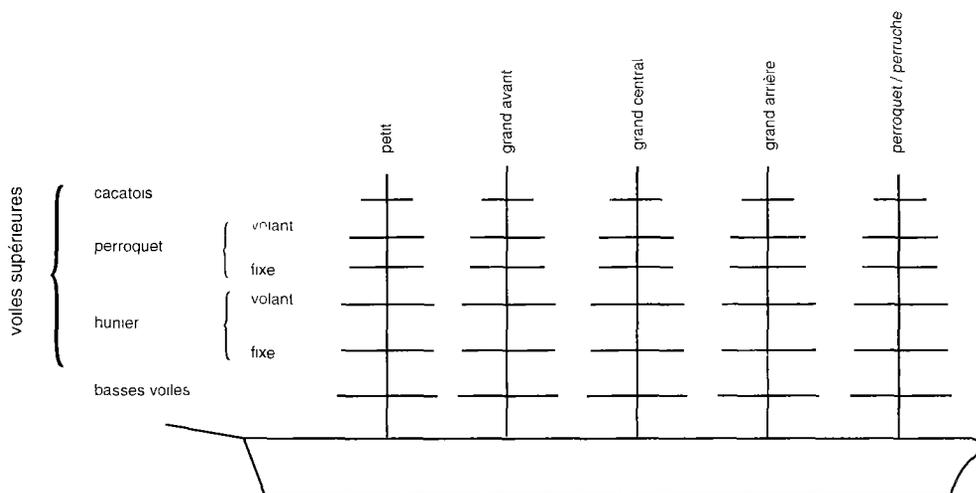


TABLEAU 1

La typologie des voiles carrées regroupe quelque 90 notions dans notre corpus (Paasch, 1901 : 338-342). Ce nombre étant beaucoup trop important, nous avons choisi de restreindre l'objet de notre démonstration aux seules voiles dénommées *cacatois* (*Royal*, en anglais et en allemand)⁹. Un décompte très précis permet de dénombrer 8 notions se rapportant aux cacatois dans le corpus. Mais ces 8 notions appartiennent à un R.N.I. trilingue et constituent le résultat de la confrontation des réseaux notionnels anglais, français et allemand. En effet, si l'on se fonde sur les légendes des illustrations et les systèmes de désignation propres à ces trois langues, on obtient des arborescences distinctes, comportant chacune un nombre différent de notions qui se rapportent pourtant toutes aux mêmes réalités matérielles. Les arborescences anglaise et allemande englobent chacune 6 notions (tableaux 2 et 3), alors que l'arborescence française en comporte 7 (tableau 4). La confrontation des langues et la prise en compte des faits de chevauchement montre que l'on aboutit au total à 10 notions différentes¹⁰.

⁹ Notre choix s'est porté sur ces voiles, car elles ne se subdivisent d'ordinaire pas en cacatois fixe et cacatois volant, ce qui a le mérite de simplifier le propos.

¹⁰ Sont 1. *Royal* = *cacatois* = *Royal*, 2. *fore-royal* = *Vor-Royal*, 3. *main-royal* = *Gross-Royal*, 4. *middle-royal* = *grand cacatois central* = *Mittel-Royal*, 5. *mizen-royal* = *Kreuz-Royal*, 6. *jigger-royal* = *Jigger-Royal*, 7. *grand cacatois*, 8. *grand cacatois avant*, 9. *grand cacatois arrière*, 10. *cacatois de perruche*.

Pourtant, pour rendre compte de la même réalité et permettre une traduction qui fonctionne quelle que soit la langue source et la langue cible, Paasch bâtit un réseau notionnel unique (R.N.I.) de 8 notions. Nous nous attacherons à découvrir dans les pages qui suivent comment une telle réduction peut se justifier.

[1] Royal	Cacatois	Royal; Oberbramsegel	
[2] Fore-royal	Petit cacatois	Vor-Royal	
[3] Main-royal	Grand cacatois	Gross-Royal	
[4] Main-royal	Grand cacatois avant	Gross-Royal	4MC 4MB 5MB ¹¹
[5] Middle-royal	Grand cacatois central	Mittel-Royal	5MB
[6] Mizén-royal	Grand cacatois arrière	Kreuz-Royal	4MC 4MB 5MB
[7] Mizén-royal	Cacatois de perruche	Kreuz-Royal	3MC
[8] Jigger-royal	Cacatois de perruche	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 341)

— relation hyponymique TY

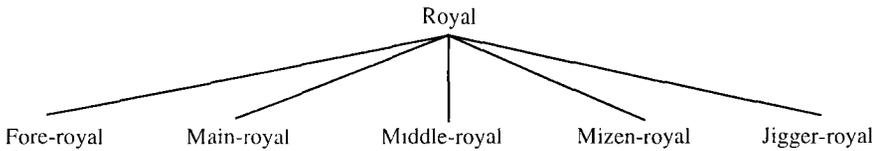


TABLEAU 2

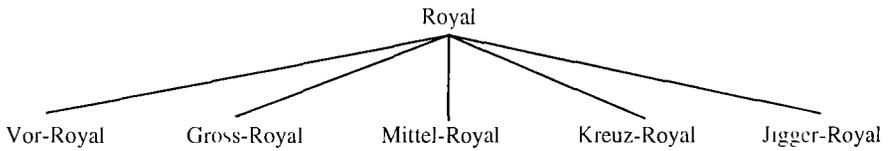


TABLEAU 3

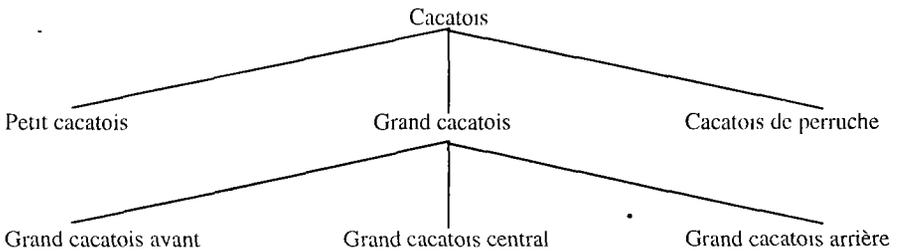
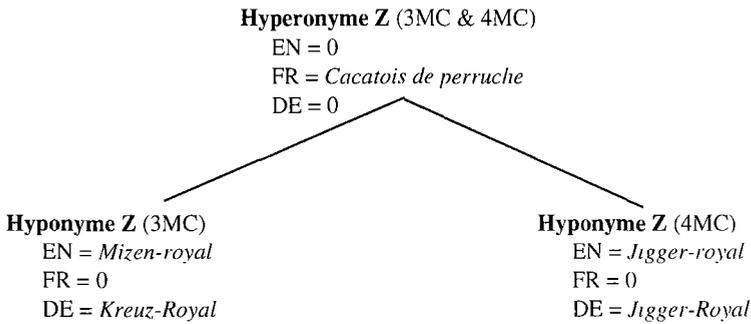


TABLEAU 4

¹¹ 3MC = trois-mâts carré, 3MB = trois-mâts barque, 4MC = quatre-mâts carré, 4MB = quatre-mâts barque, 5MB = cinq-mâts barque. Les gréements carrés se distinguent des gréements de barque par la présence de voiles carrées sur le dernier mât

3.2. Hypothèse de la notion « zéro »

Comme on le constate dans les arborescences 2, 3 et 4, les réseaux notionnels allemand et anglais sont identiques, mais différent de celui du français. Lorsqu'on fusionne ces trois arborescences dans un R.N.I. trilingue, on s'aperçoit que telle notion de telle langue ne possède pas de correspondant dans une autre langue. Ainsi, le cacatois situé tout à l'arrière d'un trois-mâts carré (3MC) ou d'un quatre-mâts carré (4MC) se nomme toujours *cacatois de perruche* en français. Par contre, les locuteurs anglophones et germanophones distinguent le cacatois de perruche d'un 3MC (*mizen-royal* = *Kreuz-Royal*) de celui d'un 4MC (*jigger-royal* = *Jigger-Royal*). La notion française *cacatois de perruche* possède donc une acception plus large et ne possède pas d'équivalents dans les deux autres langues. Inversement, les notions *mizen-royal* = *Kreuz-Royal* et *jigger-royal* = *Jigger-Royal* sont plus restreintes et ne possèdent pas d'équivalent en français. L'arborescence n° 5 confirme cette différence, qui nous conduit à distinguer trois notions au sein du R.N.I. : la notion française, perçue comme hyperonyme, et les deux notions « anglo-allemandes », perçues comme hyponymes.



TABEAU 5 : Mise en commun des trois réseaux notionnels au sein du R.N.I.

Nous proposons de désigner sous le nom de **notion « zéro »** (ci-après abrégée **notion Z**) toute notion du R.N.I. qui apparaît comme non prise en compte dans une langue précise lors de la comparaison des réseaux notionnels propres à chaque idiome.

3.3. Désignation par « hyperonomase »

Si l'on observe les équivalences proposées dans le corpus de référence, on s'aperçoit que le terminographe fournit un équivalent à la notion Z hyponyme en ayant recours à son hyperonyme immédiat¹². Ainsi, pour désigner en français les notions hyponymes *mizen-royal* = *Kreuz-Royal* et *jigger-royal* = *Jigger-Royal*, il réutilise le terme hyperonyme *cacatois de perruche*, qui, en français, renvoie à ce type de voile quel que soit le nombre de mâts.

12. Selon un principe de substitution très fréquemment attesté dans les textes spécialisés (« le *humier* » utilisé pour « le *rand humier fixe* »). Malheureusement, ce procédé fonctionne mal dans les phrases négatives (l'énoncé « *ce n'est pas un grand humier fixe* » ne peut pas toujours être remplacé par « *ce n'est pas un humier* »)

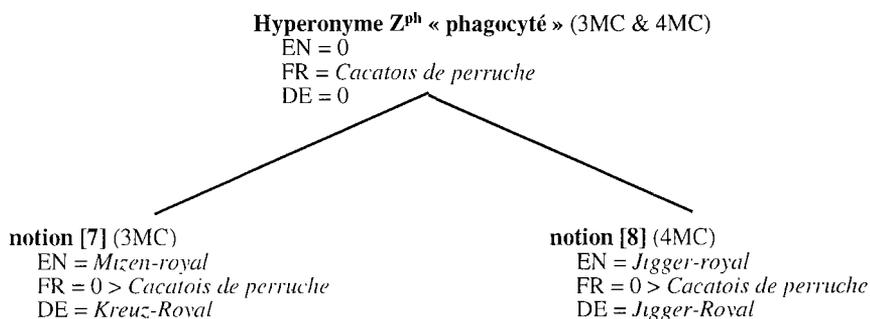


TABLEAU 6 : R N I. adapté aux besoins de la traduction

Nous risquons le terme **hyperonomase**¹³ pour rendre compte du processus qui consiste à désigner, dans une langue déterminée, une notion Z hyponyme à l'aide de son hyperonyme. Dès à présent, on perçoit que c'est l'hyperonomase qui engendre l'homonymie et que les notions Z désignées par hyperonomase ont toujours une extension plus restreinte que celle de leur hyperonyme, dont elles sont évidemment homonymes.

3.4. « Phagocytose » de l'hyperonyme Z

L'explication de l'équivalence n'est assurément pas aussi simple, car si l'hyperonomase permet de désigner les notions [7] et [8] dans chacune des langues, elle ne rend pas compte de la disparition de la notion Z hyperonyme *cacatois de perruche* dans le dictionnaire de Paasch. En effet, l'hyperonomase fournit un équivalent français aux deux notions hyponymes, mais point d'équivalents anglais et allemand à la notion hyperonyme. Dans le dictionnaire, les équivalents sont bel et bien prévus pour les notions [7] et [8], mais non pour la notion Z hyperonyme. En effet, la notion *cacatois de perruche* constitue au sein du R.N.I. un générique inutile, qui peut être aisément supprimé. Un cacatois de perruche se situe nécessairement à bord d'un 3MC (notion [7]) ou d'un 4MC (notion [8]).

Il semble bien que dans certains cas l'hyperonomase entraîne la disparition pure et simple de la notion Z hyperonyme. Elle est littéralement « phagocytée » par les notions Z hyponymes dès lors que l'hyperonomase rend inutile sa traduction dans les autres langues (on parlera ci-après de **notion hyperonyme Z phagocytée** ou **Z^{ph}**). Nous sommes persuadé que le dictionnaire de Paasch, conçu pour la traduction, obéit à ce principe de la **phagocytose**, lequel mérite assurément d'être approfondi d'un point de vue théorique.

Cette disparition pose inévitablement un problème de traduction : s'il est aisé de traduire les termes anglais *mizen-royal* et *jigger-royal* vers le français en usant de l'hyperonomase, force est de reconnaître que l'inverse n'est pas exact. Si, dans le

¹³ Ce néologisme est, certes, critiquable, mais permet d'éviter de lourdes circonlocutions. En récupérant *onomase* pour lui adjoindre *hyper-*, nous complétons la famille *hyponyme*, *hyponymie*, *hyperonyme* (proposée par Lyons, 1970 : 347) tout en suivant – du moins en synchronie – le modèle de la famille *paronyme*, *paronymie*, *paronomase*.

cadre d'une relation générique (TY), l'hyperonyme peut toujours désigner l'hyponyme (car il l'englobe), inversement, l'hyponyme ne permet pas de désigner l'hyperonyme (car il est plus restreint). Confronté au terme français *cacatois de perruche* utilisé comme générique, le traducteur hésitera entre les deux hyponymes anglais (regroupés dans le dictionnaire de Paasch) et sans doute les coordonnera-t-il dans sa traduction.

3.5. Approche théorique de l'hyperonomase et de la phagocytose

3.5.1. Le chevauchement culturel et le référent commun

Le caractère spécialisé du domaine abordé peut donner une complexité apparente aux deux phénomènes qui viennent d'être présentés : l'hyperonomase et la phagocytose. Il ne s'agit pourtant, *a priori*, que de cas où l'absence d'isomorphisme se traduit par une inclusion de la notion d'une langue dans la notion d'une autre langue.

On constate, en effet, que dans tous les cas où la phagocytose est envisageable, le référent des notions hyponymes peut être désigné par le terme hyperonyme. Dans tous les cas de notion Z^{ph} rencontrés, il apparaît que l'extension de Z^{ph} correspond parfaitement à l'addition des extensions de chacun de ses hyponymes. On dira que l'hyperonyme Z^{ph} est capable de désigner les objets conceptualisés comme co-hyponymes dans d'autres langues : il désigne les mêmes référents. Lyons (1970 : 349-350) a déjà évoqué à sa manière le problème de la notion Z^{ph} en montrant bien que dans le cadre d'une relation hiérarchique, le choix d'utiliser le terme hyperonyme pour désigner l'hyponyme permet de résoudre ce qu'il dénomme *le non-isomorphisme des langues*. Remarquons toutefois que Lyons concluait à l'absence de règle sémantique et au règne de l'intuition, constat que nous allons tenter de dépasser dans les pages qui suivent.

3.5.2. Le R.N.I. pour contexte

Il convient de rappeler que les notions Z et Z^{ph} n'existent que dans le cadre du R.N.I., c.-à-d. dans le cadre d'une confrontation des langues. À notre connaissance, le principe du recours à l'hyperonyme n'a jamais été établi en termes d'adaptation du R.N.I. aux besoins de la traduction. En accordant une si grande importance à la relation espèce-genre, vue comme foncièrement hiérarchique, Wüster avait assurément l'intuition de ce principe de l'hyperonomase. Toutefois, il n'a pas cherché à l'expliquer et ne l'a guère exploité¹⁴, dans la mesure où il acceptait difficilement l'homonymie entre l'hyperonyme et l'hyponyme, perçue comme un sommet de l'ambiguïté plutôt que comme le résultat inévitable de la confrontation des langues.

Certains termes sont tellement ambigus qu'ils désignent à la fois une notion et l'un des spécifiques de cette notion (*homonymes verticaux*). En terminologie, on distingue ces deux notions en ajoutant le chiffre romain ¹, en exposant, après

¹⁴ Dans le *Dictionnaire multilingue de la machine-outil*, Wüster (1968) utilise divers symboles qui permettent d'annoncer les cas d'équivalence partielle, mais non de les résoudre. Il ne recourt qu'exceptionnellement aux regroupements homonymiques.

le terme lorsqu'on veut parler du sens large d'un point de vue logique. Dans l'autre cas, on utilise le chiffre ¹⁵, en exposant, après le terme [...]. (Wüster, 1981 : 88)

Felber (1987 : 153) montre lui-même que ces « homonymes verticaux », qu'il nomme *homonymes polysèmes*, entretiennent bien une relation hyponymique, voire une relation partie-tout. Il ne semble toutefois pas établir de lien entre cette perspective de passage de la polysémie à l'homonymie et la comparaison des réseaux notionnels de chaque langue.

3.5.3. *Hyponyme = hyperonyme + actualisation d'un caractère virtuel*

Dans la théorie viennoise, les notions se composent d'un ensemble de **caractères**. Ces caractères, qui sont des propriétés des objets conceptualisés, permettent de différencier ou de rapprocher les notions¹⁵. Normalement, les genres sont distingués des espèces selon un même **type de caractère**, c.-à-d. en fonction de caractères fondés sur un même **critère de subdivision**¹⁶ (p. ex. le nombre de mâts, dans la subdivision des voiliers en trois-mâts, quatre-mâts, cinq-mâts, etc.).

Tous les co-hyponymes d'un même hyperonyme possèdent inévitablement un certain nombre de caractères en commun, lesquels correspondent exactement aux caractères de leur hyperonyme. Tel est par exemple le cas pour les types de grands cacatois. On constate clairement dans le tableau qui suit que les trois hyponymes *grand cacatois avant*, *grand cacatois central* et *grand cacatois arrière* possèdent les mêmes caractères que leur hyperonyme *grand cacatois*, dont ils se différencient par un caractère au moins¹⁷. Rien n'interdit toutefois de dire que l'hyperonyme possède également ces caractères de manière virtuelle. Comment expliquer autrement que le terme *grand cacatois* puisse servir à désigner chacun des trois hyponymes ? L'idée d'une prise en compte de caractères virtuels paraît d'autant plus envisageable que les terminologies dénomment fréquemment les hyponymes par des syntagmes qui adjoignent un caractère lexicalisé derrière le terme hyperonyme.

<i>grand cacatois</i>	'cacatois' ¹⁸	'sur un grand mât'	0
<i>grand cacatois avant</i>	'cacatois'	'sur un grand mât'	'avant'
<i>grand cacatois central</i>	'cacatois'	'sur un grand mât'	'central'
<i>grand cacatois arrière</i>	'cacatois'	'sur un grand mât'	'arrière'

TABLEAU 7

Nous nommerons **caractères virtuels** les propriétés d'un objet qui, dans une langue donnée, ne sont pas conceptualisées pour délimiter la notion alors qu'elles le sont

15 « Caractère · Représentation mentale d'une propriété d'un objet (2.1) et qui sert à en délimiter la notion (3.1). » (ISO 1087, 1990 : 2)

16 « Type de caractère · Toute catégorie de caractère utilisée comme critère dans l'établissement d'un système de notions générique » (ISO 1087, 1990 : 2) La norme ISO 704 (1987 : 4) parle de *critère de subdivision*, terme qui nous paraît plus transparent et que nous adopterons dans la suite de l'étude.

17 « [...] le concept spécifique a les caractères du concept générique plus un au moins. Au fur et à mesure qu'on monte vers du plus général, on est en présence de concepts dits plus 'abstrait' » (Lerat, 1990 : 81).

18. Par convention, les caractères sont représentés entre guillemets simples.

dans d'autres. Une telle approche implique, pour la rigueur du propos, que l'on reconsidère la définition de la notion proposée par l'ISO 1087 (1990 : 1) : dans un contexte multilingue, la notion doit, en effet, être définie comme la conceptualisation d'un ou de plusieurs objets à partir de certaines de leurs propriétés (caractères), identifiées comme pertinentes dans une langue donnée.

3.5.4. R.N.I. et instabilité notionnelle

En terminologie, les notions sont réputées stables. Ce qui est vrai tant qu'on demeure dans une perspective monolingue tend pourtant à devenir très relatif dans une perspective multilingue. En effet, dans le cadre d'une recherche d'équivalences au sein du R.N.I., le terme n'apparaît plus comme monosémique et, selon son sens, se traduira de telle ou telle manière. Ainsi, confronté à l'anglais et à l'allemand, le sens du terme français *cacatois de perruche* se met à varier (la notion se dédouble). Le tableau 8 mentionne les caractères considérés comme pertinents dans le système notionnel de chaque langue pour distinguer les cacatois de perruche. On y observe une correspondance exacte des caractères de la notion anglaise *mizen-royal* avec ceux de la notion allemande *Kreuz-Royal*, d'une part, et des caractères de la notion anglaise *Jigger-royal* avec ceux de la notion allemande *Jigger-Royal*, d'autre part.

FR : <i>cacatois de perruche</i> ¹⁰	:	'cacatois'	'dernier mât'	0
EN : <i>mizen-royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 3MC'
EN : <i>jigger-royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 4MC'
DE : <i>Kreuz-Royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 3MC'
DE : <i>Jigger-Royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 4MC'

TABLEAU 8

On remarquera également que les caractères « à bord d'un 3MC » et « à bord d'un 4MC » ne sont pas nécessaires pour décrire la notion française *cacatois de perruche*. Toutefois, ces mêmes caractères « à bord d'un 3MC » et « à bord d'un 4MC », deviennent pertinents dans le cadre d'une traduction vers l'anglais ou l'allemand. En effet, dans le cadre du R.N.I., ces caractères doivent être pris en considération de manière à trouver une équivalence en vertu du principe d'équivalence notionnelle ; c.-à-d. que pour arriver à désigner le même objet, il apparaît indispensable de le conceptualiser de la même manière.

FR : <i>cacatois de perruche</i> ¹¹	:	'cacatois'	'dernier mât'	0 > 'à bord d'un 3MC'
EN : <i>mizen-royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 3MC'
DE : <i>Kreuz-Royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 3MC'
FR : <i>cacatois de perruche</i> ¹²	:	'cacatois'	'dernier mât'	0 > 'à bord d'un 4MC'
EN : <i>jigger-royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 4MC'
DE : <i>Jigger-Royal</i>	:	'cacatois'	'dernier mât'	'à bord d'un 4MC'

TABLEAU 9

En théorie, tout hyperonyme, même éloigné, est apte à désigner de manière univoque le même objet que son lointain hyponyme : il suffit que la notion hyperonyme ne se distingue de la notion hyponyme que par des caractères virtuels. Cruse (1986 : 155) fait toutefois remarquer que le recours à l'hyperonyme engendre toujours une sous-spécification. Ceci explique sans doute que Paasch utilise toujours l'hyperonyme immédiat, lequel est d'ailleurs perçu comme le plus utile par Pierre Lerat (1988 : 20).

3.5.5. Pourquoi la phagocytose ?

Dans le corpus de référence, seules les deux notions hyponymes subsistent après phagocytose de la notion Z^{ph} *cacatois de perruche*^[0]. Par le mécanisme de l'hyperonymase, les caractères virtuels de l'hyperonyme sont activés au niveau hyponymique, de sorte qu'il se révèle apte à désigner chaque hyponyme. Dans la mesure où tous les caractères virtuels de l'hyperonyme Z^{ph} se trouvent ainsi activés sous toutes leurs valeurs possibles au niveau des hyponymes, ledit hyperonyme Z^{ph} ne désigne plus aucun objet qui ne soit concrètement représenté par ses hyponymes. La notion hyperonyme Z^{ph} devient donc superflue au sein du R.N.I. d'un dictionnaire de traduction et peut être phagocytée.

[7] Mizen-royal	Cacatois de perruche	Kreuz-Royal	3MC
[8] Jigger-royal	Cacatois de perruche	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 : 341)

Dans le cadre d'une entreprise visant à permettre la communication entre des locuteurs de langues différentes, il paraît plus utile de traduire deux notions spécifiques dans la langue qui ne les désignait pas que de rendre compte d'une notion générique que ladite langue était la seule à prévoir. S'agissant de désigner des objets, l'extension du générique correspond toujours au total des extensions des notions spécifiques. Dès lors que ces dernières sont dénommées dans chacune des langues, le générique ne constitue plus qu'une abstraction de peu d'utilité. La phagocytose paraît ainsi s'imposer d'elle-même dans le cas de la distinction entre les notions *mizen-royal = cacatois de perruche*^[1] = *Kreuz-Royal* (3MC) et *jigger-royal = cacatois de perruche*^[2] = *Jigger-Royal* (4MC), lesquelles rendent inutile toute référence à une notion recouvrant en même temps le cacatois de perruche d'un 3MC et celui d'un 4MC. En d'autres termes, au sein du R.N.I., l'hyperonyme Z^{ph} ne désigne rien que ne désignent déjà ses hyponymes.

L'actualisation du caractère virtuel au niveau hyponymique entraîne en traduction une modification implicite du sens de l'hyperonyme, mais il ne s'agit jamais que d'un artifice terminographique visant à établir l'équivalence. Que l'on parle du cacatois de perruche d'un 3MC ou de celui d'un 4MC, on le désigne toujours par le terme *cacatois de perruche*. Jamais il n'est demandé au locuteur francophone de revoir la manière dont il appréhende le réel au nom d'une quelconque normalisation.

Par ailleurs, si les notions sont classées en vertu du lien TY – et tel est le cas chez Paasch –, les homonymes nés d'une phagocytose se retrouvent normalement regroupés dans le dictionnaire. Le traducteur francophone peut ainsi découvrir que dans le cadre d'un contexte anglais ou allemand qui établit une nette distinction entre

mizen-royal et *jigger-royal*, entre *Kreuz-Royal* et *Jigger-Royal* (cf. 3.4.), il convient de spécifier davantage la portée du terme *cacatois de perruche* en y adjoignant un complément déterminatif (*de trois-mâts carré, de quatre-mâts carré*). Réciproquement, un traducteur anglais ou allemand découvrira que la traduction du générique français *cacatois de perruche* appelle une interprétation du contexte pour décider du caractère virtuel activé.

3.5.6. Existe-t-il des hypernomases sans phagocytose ?

Dans les exemples produits jusqu'à présent, l'hypernomase s'accompagne toujours d'une phagocytose de l'hyperonyme Z. Il est assurément des cas où l'hyperonyme n'est pas une notion Z et n'est donc pas « phagocytale ». Toutefois, ces cas sont rares dans *De la quille à la pomme de mât*, sauf exceptions propres au chapitre des *Termes généraux*. C'est ainsi que la possibilité d'opérer une distinction entre les notions *observatoire astronomique* et *observatoire météorologique* n'exclut pas la nécessité de devoir éventuellement faire référence à la notion générique *observatoire*, dans le cas d'un établissement qui réunirait les deux fonctions, voire davantage.

Observatory. Any place from where a view may be observed.

Point d'observation. Un endroit quelconque duquel on jouit d'une vue

Warte. Ein erhabener Ort von wo man eine freie Aussicht hat

Observatory. A building fitted with installations and instruments necessary for making astronomical, etc observations.

Observatoire. Etablissement pourvu des installations et des instruments nécessaires pour les observations astronomiques, météorologiques, etc.

Warte. Ein für astronomische, meteorologische u s.w. Beobachtungen eingerichtetes, und mit den hierzu erforderlichen Instrumenten ausgestattetes Institut.

Observatory. (astronomical)

Observatoire (astronomique)

Sternwarte.

Meteorological-observatory.

Observatoire (météorologique)

Wetterwarte.

(Paasch 1901 : 505)

EN = *Observatory*
FR = *Observatoire*
DE = *Warte*

EN = 0 > *Observatory*
FR = 0 > *Observatoire*
DE = *Sternwarte*

EN = *Meteorological-observatory*
FR = 0 > *Observatoire*
DE = *Weiterwarte*

TABEAU 10

3.5.7. La relation TY, condition de l'hyperonomase et de la phagocytose

Comme on le constate, l'arborescence n° 10 ne peut rendre compte de l'équivalence *observatory* = *point d'observation* = *Warte*, car la notion ainsi désignée n'appartient pas à la même arborescence TY. Il nous semble important de remarquer que si l'hyperonomase répond au principe d'équivalence notionnelle, elle n'en est qu'une forme d'accomplissement très particulière, liée à une relation hyponymique observable au sein du R.N.I.

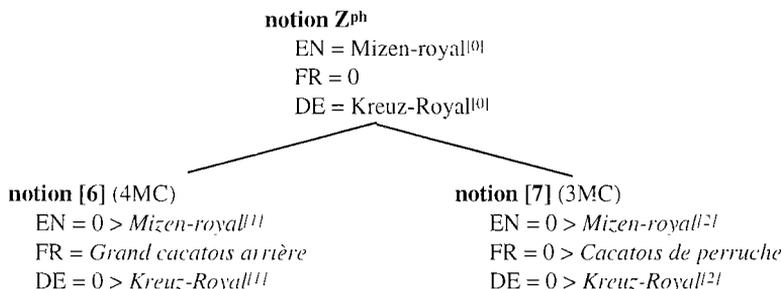
Il ne peut être question de parler d'hyperonomase et encore moins de phagocytose pour les cas cités en 2.3 (*watch*, *pilotage*, *route*), car aucune des trois langues ne possède une notion qui serait l'hyperonyme de notions propres aux autres langues. Les caractères communs permettent tout au plus de déterminer un lien notionnel indéterminé qui fonde une relation fonctionnelle. Par exemple, la bordée (*watch*) « est responsable de » la veille (*watch*) « pendant » le quart (*watch*) ; ou encore, le bureau de pilotage (*pilotage*) « est le centre des activités de » pilotage (*pilotage*).

3.6. Équivalence partielle et hyponyme virtuel Z'

La confrontation des notions peut aboutir à observer des cas où deux langues ne possèdent pas de désignation pour une réalité particulière, conçue comme incluse dans des notions plus larges mais dont l'extension varie d'une langue à l'autre. L'étude du dictionnaire de Paasch nous conduit à poser l'hypothèse de l'existence dans le R.N.I. de **notions zéros virtuelles (Zv)**. Il s'agit de notions Z hyponymes qui, bien qu'elles ne soient propres à aucune langue, doivent être désignées par hyperonomase pour résoudre le problème d'équivalence posé par de tels cas.

3.6.1. Un cas de notion Z'

Le phénomène de la notion Z' s'observe dans l'arborescence des types de cacatois. Les notions équivalentes *mizen-royal* = *Kreuz-Royal* correspondent à deux notions hyponymes en français : *grand cacatois arrière* (à bord d'un 4M ou d'un 5M) et *cacatois de perruche* (à bord d'un 3M). Conformément au principe de l'hyperonomase et de la phagocytose, le dictionnaire ne retient donc que les deux notions hyponymes du R.N.I. : d'une part, *Mizen-royal*^[1] = *Grand cacatois arrière* = *Kreuz-Royal*^[1] et, d'autre-part, *Mizen-royal*^[2] = *Cacatois de perruche* = *Kreuz-Royal*^[2].



TABEAU 11

Si l'on considère à présent l'ensemble des désignations des cacatois au sein du R.N.I., on s'aperçoit que le second hyponyme (notion [7]) correspond à une notion que nous avons déjà décrite comme résultant d'une autre hyperonomase accompagnée de phagocytose, celle décrite dans le tableau 6. Dans la mesure où tout ceci doit paraître bien abstrait, nous avons essayé de recréer, dans le tableau 11 une vue d'ensemble du R.N.I. avant phagocytose. La partie gauche de l'arborescence correspond au tableau 11 ; la partie droite, au tableau 6. Pour clarifier les notions, nous avons représenté les objets conceptualisés (voiles) par chacune d'entre elles. Il apparaît clairement que la notion [7] conceptualise exactement le même objet, quand bien même elle peut être appréhendée au départ d'hyperonymes distincts, mais possédant des caractères parfaitement compatibles : la partie gauche de l'arborescence distingue deux types de cacatois en fonction de l'emplacement du mât ; la partie droite, en fonction du nombre de mâts. Ceci explique que dans le R.N.I. de *De la quille à la pomme de mât*, il ne s'agit que d'une seule et même notion, celle que nous avons identifiée par le chiffre [7].

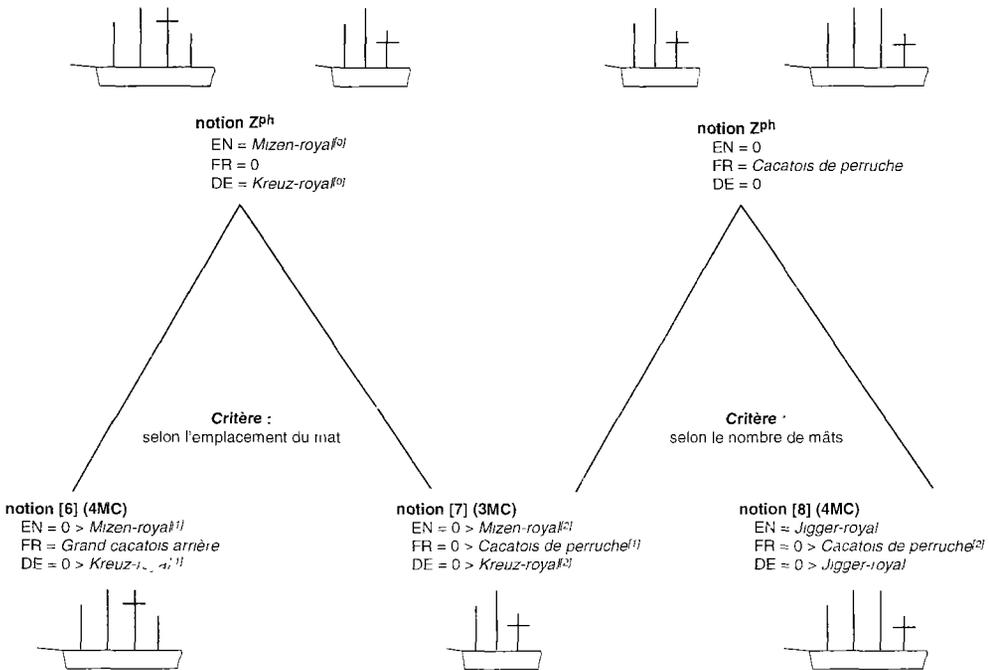


TABLEAU 12

[6] Mizen-royal ¹¹	Grand cacatois arrière	Kreuz-Royal ¹¹	4MC, 4MB, 5MB
[7] Mizen-royal ²¹	Cacatois de perruchel ¹¹	Kreuz-Royal ¹²	3MC
[8] Jigger-royal	Cacatois de perruchel ²¹	Jigger-Royal; Besahn-Royal	4MC

(Paasch 1901 : 341)

Il s'agit d'un cas patent de notion Z^v. En effet, la reconstitution du R.N.I. montre que la notion [7] n'existe dans aucune langue : elle est tout à la fois hyponyme de la

notion Z^{ph} *Mizen-royal*^[0] = *Kreuz-Royal*^[0] (tableau 11) et de la notion Z^{ph} *cacatois de perruche*^[0] (cf. tableau 6). La notion interlinguistique [7], présente dans le dictionnaire, est clairement une notion qui n'existe ni en anglais, ni en français, ni en allemand. Aucune de ces trois langues ne possède une notion aussi restreinte, qui ne renverrait qu'au seul cacatois du dernier mât d'un trois-mâts carré (3MC). Si le terminographe a créé cette notion virtuelle « de toute pièce », c'est bien pour permettre la traduction la plus juste, compte tenu de tous les référents envisageables.

L'illustration et la dénomination des référents permettent d'ailleurs d'aboutir empiriquement à une solution rigoureusement identique. Dans le tableau 13, inspiré du problème du découpage des couleurs proposé par Lyons (1970 : 46-47)¹⁹, chaque case correspond à chacun des objets (cacatois) désignés dans chaque langue par un terme différent ; en d'autres termes, chaque case représente l'extension de la notion dénommée par ce terme. La confrontation des découpages confirme bien que dans le cadre d'un dictionnaire trilingue, il faut envisager trois notions différentes au sein du R.N.I. pour arriver à désigner les trois référents envisageables.

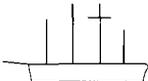
 FR = Grand cacatois arrière	 FR = Cacatois de perruche	
 EN = Mizen-royal	 EN = Jigger Royal	
 DE = Kreuz-Royal	 DE = Jigger-Royal	
notion [6] (4MC) FR = Grand cacatois arrière EN = Mizen-royal DE = Kreuz-Royal	notion [7] (3MC) FR = Cacatois de perruche EN = Mizen-royal DE = Kreuz-Royal	notion [8] (4MC) FR = Cacatois de perruche EN = Jigger-royal DE = Jigger-Royal

TABLEAU 13

3.6.2. Un cas complexe et exemplaire

Le modèle des notions Z^{ph} et Z^{v} permet d'expliquer la manière dont de nombreux problèmes d'équivalence particulièrement complexes ont été résolus au sein du corpus de référence. On sera, par exemple, intrigué d'y découvrir quatre entrées *diablotin*, alors

19 Nous avons déjà analysé ailleurs le lien entre le problème des couleurs et le P.E.N. (Van Campenhoudt, 1991 et 1994 70-71)

que les marins français considèrent que ce terme désigne une seule et même voile triangulaire (voile d'étai), toujours située devant le dernier mât (le mât d'artimon).

[1] Mizen-topmast-staysail	Diablotin	Kreuz-Stengestagsegel	(3MC)
[2] Mizen-topmast-staysail	Diablotin	Besahn-Stengestagsegel	(3MB, BAR, 3MG)
[3] Jigger-topmast-staysail	Diablotin	[Kreuz-Stengestagsegel] ²⁰	(4MC)
[4] Jigger-topmast-staysail	Diablotin	Besahn-Stengestagsegel	(4MB, 5MB)

(Paasch 1901 : 343)

Un examen approfondi montre que l'on distingue en allemand deux types de diablotin selon que cette voile se situe devant un mât d'artimon qui ne comporte que des voiles axiales (syntagme formé avec *Besahn*) ou qui comporte également des voiles carrées (syntagme formé avec *Kreuz*). En anglais, on se fonde sur le nombre de mâts pour distinguer les diablotins d'un 3M (syntagme formé avec *mizen*) et ceux d'un 4M (syntagme formé avec *jigger*). En français, on considère qu'il s'agit à chaque fois d'une seule et même voile. Le tableau 14 (arborescence) rend compte de la situation de ces notions au sein du R.N.I. dès lors que l'on prend en compte les caractères distinctifs propres à chacune des langues considérées.

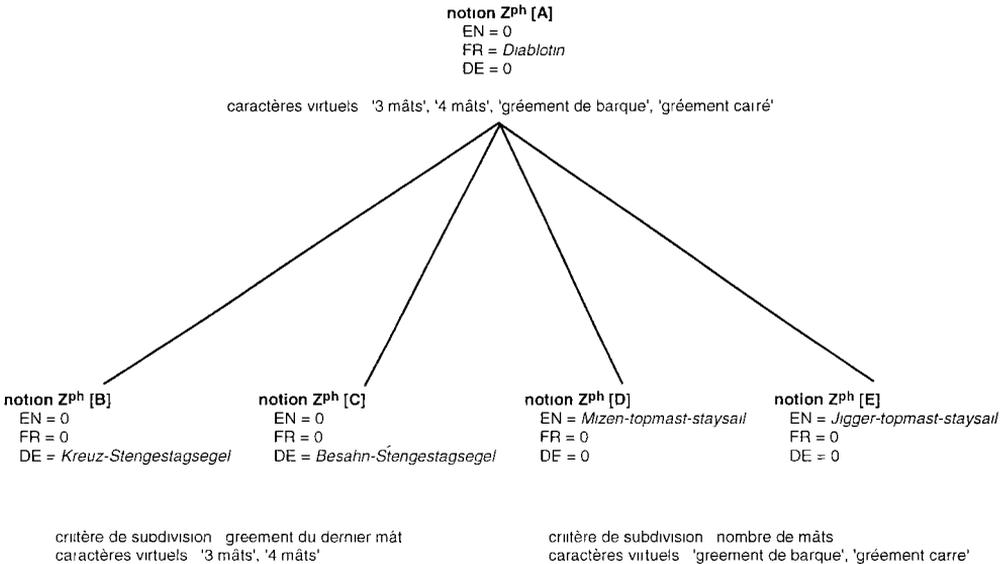


TABLEAU 14

La solution du dictionnaire correspond à la prise en compte de notions virtuelles Z^y qui préservent l'intégrité référentielle de chaque terme tout en permettant la traduction dans un R.N.I. trilingue. Cette solution est représentée sous forme d'arborescence dans le tableau 15. On y découvre que l'extension très large de la notion *diablotin* en

20 Par souci de simplifier l'exposé, nous reproduisons exceptionnellement le terme allemand tel qu'il apparaît dans la quatrième édition du dictionnaire (1908), après stabilisation du système de désignation

français fait de celle-ci un hyperonyme Z^{ph} , tant vis-à-vis des notions anglaises que vis-à-vis des notions allemandes. Le découpage hyponymique en fonction de la disposition des voiles (en allemand) ou du nombre de mâts (en anglais) intervient au niveau immédiatement subordonné. On considérera donc que dans le R.N.I., il existe quatre hyponymes de *diablotin*^[0] : *Kreuz-Stengestagssegel*^[0], *Besahn-Stengestagssegel*^[0], *Mizen-topmast-staysail*^[0] et *Jigger-topmast-staysail*^[0].

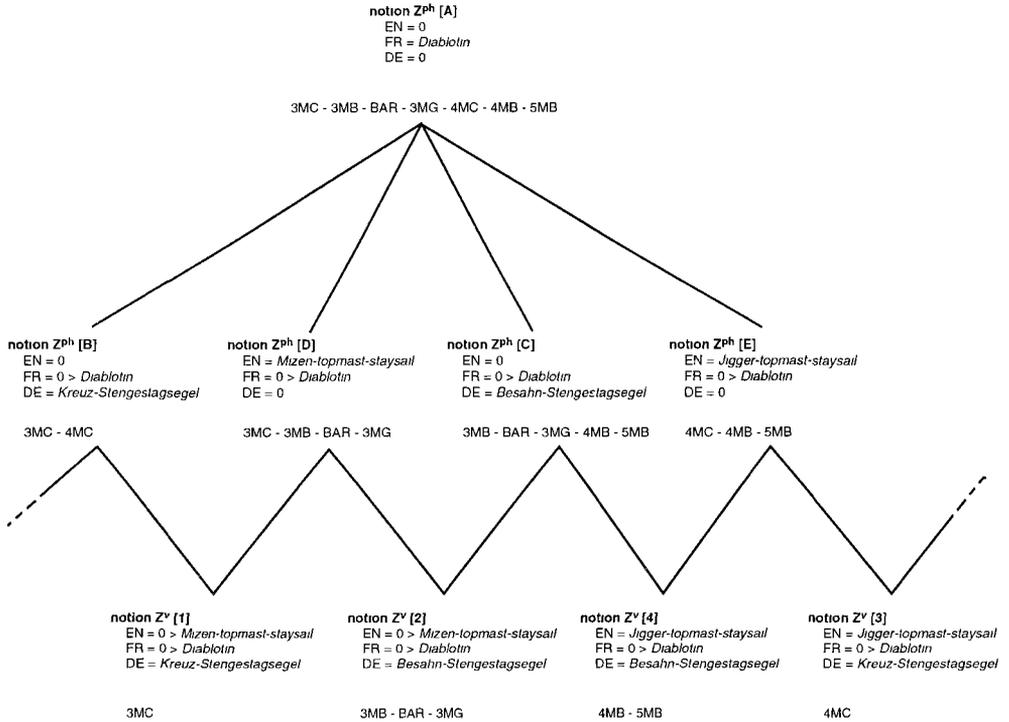


TABLEAU 15

Paradoxalement, toutes ces notions co-hyponymes constituent des notions Z^{ph} , à l'instar de *diablotin*^[0]. En effet, elles ne possèdent aucun équivalent dans les deux autres langues. Toutefois, l'activation des caractères virtuels – qui correspond, plus simplement, à la prise en considération des référents – permet de dégager quatre notions Z^v , propres à aucune langue, mais aptes à permettre une traduction dans les six sens envisageables dès lors qu'on les désigne au moyen de leurs hyperonymes respectifs. On observe dans l'arborescence n° 15 que les quatre notions Z^v correspondent parfaitement aux notions mentionnées et illustrées dans le dictionnaire.

3.7. Approche théorique de la notion virtuelle

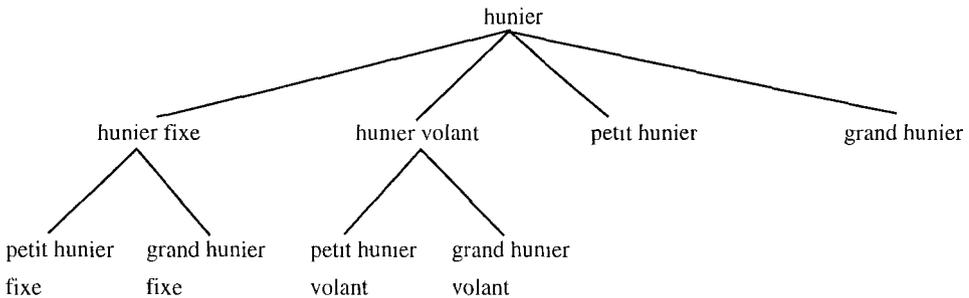
Ce mode de résolution suppose que dans un réseau multilingue une notion hyponyme puisse être subordonnée à deux hyperonymes Z^{ph} différents. En fait, même dans un réseau monolingue, une notion hyponyme peut dépendre de plusieurs hyperonymes dès lors qu'elle conserve en les combinant les caractères différenciateurs desdits hypero-

nymes et qu'elle actualise leurs caractères virtuels. Une telle notion ne possédera pas de caractère différenciateur propre.

Force est de constater que dans les cas rencontrés, la notion virtuelle Z^v combine toujours des caractères propres au système hyponymique de chacune des langues prises en compte. Les caractères combinés constituent ce que la théorie viennoise nomme des **caractères indépendants**²¹.

Les caractères sont dits **dépendants** lorsqu'ils doivent nécessairement intervenir à des niveaux hyponymiques différents de la hiérarchie arborescente (Felber 1987 : 100). Par exemple, dans la distinction des types de navires, le caractère « muni d'une chaudière » précède nécessairement des caractères comme « muni de roues à aubes » ou « muni d'une hélice », auxquels il est supérieur. En effet, les notions *vapeur à roues* et *bateau à vapeur à hélice* sont hiérarchiquement subordonnées à la notion (*bateau à*) *vapeur*.

Par contre, les caractères **indépendants** « peuvent se suivre à différents niveaux d'une série verticale de notions et être combinés arbitrairement » (Felber, 1987 : 101). En d'autres termes, dans une même arborescence TY, des caractères indépendants peuvent servir à distinguer des co-hyponymes sur la base de critères de subdivision différents. Par exemple, les caractères « fixe » ou « volant », d'une part, et « situé sur le mât d'artimon » ou « situé sur le grand mât », d'autre part, permettent de distinguer quatre types particuliers de la notion *hunier* : *hunier fixe* et *hunier volant* d'une part, *petit hunier* et *grand hunier* d'autre part. Fondés sur des critères différents (la mobilité et l'emplacement), ces caractères sont indépendants dans la mesure où ils peuvent se combiner à un niveau inférieur pour distinguer les notions *petit hunier volant*, *petit hunier fixe*, *grand hunier volant* et *grand hunier fixe*.



(d'après Paasch 1901 . 338-339)

TABLEAU 16

Les notions hyponymes situées au croisement de deux typologies peuvent avoir autant d'hyperonymes qu'elles actualisent de critères propres à chacun d'eux. Ainsi, la notion *petit hunier volant* possède deux hyperonymes (*petit hunier* et *hunier volant*) et constitue le point de liaison de deux arborescences fondées chacune sur un critère de subdivision différent : la mobilité et le mât.

21 La distinction entre caractères indépendants et caractères dépendants n'est malheureusement plus prise en compte dans les dernières normes ISO 704 (1987) et ISO 1087 (1990)

3.8. Vers une application informatique

Notre approche théorique montre que l'équivalence possède un fondement relativement logique lorsqu'elle est obtenue au sein de la relation TY. Ce constat nous conduit à penser qu'une B.C.T. multilingue devrait être à même de déceler, voire de traiter, les notions Z au sein du R.N.I.

3.8.1. Du réseau à l'équivalence

Dans la pratique, le R.N.I. d'un glossaire multilingue se doit d'être immédiatement utile pour le traducteur qui souhaite connaître l'équivalent idoine. Cette perspective est celle qui est logiquement suivie dans un dictionnaire conçu et présenté sur papier, tel que *De la quille à la pomme de mât*.

Idéalement, une B.C.T. multilingue devrait proposer un réseau par langue. Le R.N.I. ne serait constitué que dans un second temps par une comparaison des notions de chaque langue. Une exploitation logique de chaque réseau, fondée notamment sur les caractères et la relation TY, devrait permettre d'isoler les cas de non-isomorphisme et de proposer des équivalences acceptables. Jusqu'à cette date, aucun logiciel gestionnaire de données terminologiques n'a réellement été développé dans cette perspective²². Or, la construction d'un réseau notionnel interlinguistique peut se révéler complexe et risque d'être remise en cause dès qu'il sera décidé d'y intégrer une nouvelle langue.

Un logiciel de terminologie « intelligemment assistée par ordinateur » devrait, en réalité, être à même de formuler diverses propositions face aux impossibilités de traduction. Ainsi, lors d'une phase d'évaluation qui suivrait l'élaboration des réseaux de chaque langue, il pourrait émettre diverses propositions comme :

Proposition 1 : « Le terme français *cacatois de perruche* n'a pas d'équivalent en anglais, voulez-vous connaître les notions hyponymes en anglais ? »

Proposition 2 : « Les hyponymes anglais de *cacatois de perruche* sont *mizen-royal* et *jigger-royal* et n'ont pas d'équivalents en français. Voulez-vous utiliser l'hyponyme *cacatois de perruche* pour désigner ces hyponymes en français ? »

3.8.2. Implications définitoires

Toutefois, des précautions s'imposent : il ne saurait être question de permettre au logiciel d'altérer l'information initiale. Il s'agit plutôt d'exploiter celle-ci, de l'interpréter à la manière d'un véritable système expert chargé de seconder le terminologue.

L'idée d'une hypernomase et d'une phagocytose assistées demeure, bien sûr, une hypothèse qui doit se vérifier à l'épreuve des faits. Elle pose des problèmes qui méritent d'être étudiés avec beaucoup de précautions, notamment le report de la définition de l'hyponyme phagocyté au niveau de l'hyponyme. La conséquence logique de l'hypernomase serait que dans le R.N.I., l'extension des hyponymes re-

portés au niveau subordonné (p. ex. *cacatois de perruche*) soit plus restreinte que celle qu'ils auraient dans un ouvrage monolingue²³.

4. Synthèse : R.N.I. et approche notionnelle

4.1. Une approche contrastive du découpage notionnel

Il est apparu en 2.3. et 2.4. que le principe d'équivalence notionnelle conduit à une multiplication des homonymes au sein du R.N.I. La plupart des notions citées en guise d'exemples sont classées par Paasch dans le chapitre des *Termes généraux*. Elles ne sont guère spécialisées et pourraient fort bien être traitées de la même manière dans un dictionnaire de traduction consacré à la langue générale. Par exemple, pour déterminer l'équivalent anglais du mot *coque*, il faut nécessairement préciser à quel concept on entend faire référence : à l'enveloppe d'un fruit ou d'un œuf (= *shell*), à la carapace d'un mollusque (= *cockle*), à la carène d'un navire (= *hull*), etc.

En fait, il semble bien que le principe du dégroupement homonymique n'est qu'un avatar de la distinction entre homonymie et polysémie dans la tradition lexicographique. Tant que la perspective demeure monolingue, le lexicographe qui adopte une perspective homonymique ne dispose que de peu de critères pour décider s'il y a lieu ou non d'attribuer plusieurs entrées à un même signifiant. Par exemple, *pomme* reçoit quatre entrées dans le *D.F.C.* (1966) et six dans le *Lexis* (1987). Par contre, les dictionnaires fondés sur l'approche polysémique, comme les *Petit* et *Grand Robert*, utilisent le critère de l'étymologie pour décider du nombre d'entrées : ainsi, ils distinguent deux entrées *bière*, parce que le néerlandais *bier* et le francique *bera* ont connu des évolutions qui conduisent à attribuer des signifiants identiques à deux variétés bien distinctes de contenants.

Comparaison n'est certes pas raison, mais ce critère étymologique équivaut en quelque sorte, *mutatis mutandis*, à se servir de systèmes conceptuels propres à des langues étrangères pour présider au dégroupement homonymique. Traduites par exemple en anglais, les deux entrées *bière* requièrent des traductions différentes : *beer* et *coffin*, ce qui justifie l'existence de deux notions différentes au sein du R.N.I.

Les théoriciens de la lexicographie ont déjà abondamment disserté sur le caractère arbitraire du dégroupement homonymique basé sur une approche purement sémantique des notions. Il est bon de souligner que le P.E.N. fonde la norme non point sur des décisions arbitraires, mais sur une approche contrastive et une prise en compte de l'usage. Finalement, s'il y a un écart entre deux approches du sens, celui-ci sépare moins la terminographie et la lexicographie en soi que l'approche monolingue et l'approche multilingue. Cette dernière est *nécessairement* conceptuelle. Dès lors, s'il importe de faire référence à Wüster, ce n'est pas au nom d'une adhésion inconditionnelle à ses théories – vieillies sous plus d'un aspect –, mais parce que sa pensée fait écho à l'impérieuse nécessité, pour la traduction des langues de spécialité, de veiller à délimiter clairement le champ de l'équivalence.

²³ Paasch a veillé à arranger ses définitions en fonction des hyperonymes et phagocytoses qu'il a réalisées et des dégroupements homonymiques qui en découlent

4.2. Notion zéro et relation d'hyponymie

Au fondement de l'hypothèse développée dans cet article, se situe la notion zéro. Par-delà son appellation nouvelle, nous pensons que ce concept doit faire figure d'évidence aux yeux de tout terminologue attentif à la prise en compte des liens notionnels. Lyons (1970 : 348) constatait déjà que « les vocabulaires des langues naturelles ont tendance à présenter beaucoup de cases vides, d'asymétries et d'indéterminations » à la différence de ce qui se produit dans les taxinomies scientifiques. Cruse (1986 : 145ssq.) a longuement montré que, dans une perspective monolingue, la prise en compte de la relation espèce-genre conduisait à observer des « vides » (*gaps*) à divers niveaux de superordination de l'arborescence TY.

Par la confrontation des réseaux de différentes langues dans le cadre du R.N.I., notre étude confirme cette hypothèse et montre que le cas du vide notionnel peut également concerner le bas de l'arborescence. De ce point de vue, il faut admettre que la physionomie du réseau des relations hyponymiques observées en terminologie nautique demeure proche de celle observée dans la langue générale. On peut penser qu'il en va de même dans les nombreux domaines de spécialité qui possèdent une longue histoire et qui, au contraire des taxinomies visées par Lyons (*ibid.*), ne font l'objet d'aucune harmonisation interlinguistique.

Le concept de *notion zéro* ne s'applique pas seulement au cas du vide notionnel classique. En effet, le dépouillement de *De la quille à la pomme de mât* atteste l'existence de notions virtuelles Z^y , qui n'existent dans aucune langue mais qui sont nécessaires à l'établissement de l'équivalence dans le R.N.I. À notre connaissance, ce phénomène n'a jamais été décrit de la sorte²⁴. Pourtant les notions Z^y correspondent à des cas de chevauchement culturel et doivent être impérativement prises en compte si l'on veut bâtir une B.C.T. rigoureuse, apte à fournir des équivalents fiables. On peut penser, en effet, que l'oubli des notions Z^y explique un bon nombre d'insuffisances des dictionnaires lorsqu'il s'agit de résoudre des problèmes d'équivalence partielle.

4.3. L'équivalence partielle revisitée

Les principes d'exploitation des notions Z ont été dégagés dans cet article à partir de faits terminographiques concrets observés au sein de systèmes clos. Ils semblent permettre une description plus fine des problèmes de chevauchement culturel que ne le permet la classique distinction entre « supériorité » et « intersection » présentée par Felber (1987 : 129). Cette représentation paraît, en effet, peu adéquate pour rendre compte du rapport étroit entre l'équivalence et la relation TY, qui, l'une comme l'autre, sont identifiées à l'aide des caractères des notions concernées. Certes, Felber utilise les caractères pour expliquer l'équivalence, mais il n'approfondit pas la liaison entre lesdits caractères et la relation TY et néglige ainsi le rôle fréquent de la relation hyponymique dans l'établissement d'une équivalence.

²⁴ La norme ISO R 1087 (1969 · 8) précise qu'« une notion peut résulter de la combinaison d'autres notions, même sans égard pour la réalité », mais cet énoncé vise plutôt des découvertes scientifiques annoncées et non encore vérifiées.

Dans toute arborescence fondée sur la relation espèce-genre TY, il y a une intersection partielle entre les compréhensions des co-hyponymes. Cette intersection rassemble tous les caractères communs aux co-hyponymes. Toutefois, il ne faut pas nécessairement assimiler le phénomène de l'intersection de deux notions à celui de l'équivalence partielle, parfois nommée *intersection partielle*. Ce n'est pas parce que deux notions possèdent plusieurs caractères en commun que l'on peut parler d'équivalence partielle : qui songerait à évoquer une équivalence partielle entre les mots *sail* et *drap de lit* du fait qu'ils désignent des notions qui partagent les caractères « tissu », « blanc » et « couture » ?

Le fait même de parler d'*intersection* sans faire référence à la relation hyponymique apparaît donc comme gênant. Tant dans le cas du phénomène dit de la *supériorité* que dans celui dit de l'*intersection*, les caractères de la notion hyperonyme correspondent à l'intersection en compréhension des caractères des différents co-hyponymes. Dans un cas comme dans l'autre, la notion hyperonyme représente donc une possibilité de dénomination du subordonné par recours au principe de l'hyperonomase. Notre analyse de diverses équivalences (*mizen-royal*, *cacatois de perruche*) atteste d'ailleurs que l'hyperonomase fonctionne aussi bien dans le cas dit de la *supériorité* que dans celui de l'*intersection*.

Assimiler l'intersection partielle à l'équivalence partielle peut même conduire à négliger le cas des notions virtuelles. Ainsi, on pourrait être tenté de dire que la notion *Z'* *mizen-royal* = *cacatois de perruche* = *Kreuz-Royal* constitue un cas « d'intersection partielle » entre le français d'une part et l'anglais et l'allemand, d'autre part. Or, cette notion est clairement un cas de conjonction²⁵ en compréhension (c.-à-d. du point de vue des caractères concernés). Parler ici d'intersection, équivaut à considérer la notion en extension (c.-à-d. du point de vue des référents concernés) et à traiter de l'équivalence partielle en mélangeant deux approches définitoires fondamentalement différentes.

Les concepts théoriques de la notion zéro – qu'elle soit hyponyme ou hyperonyme, réelle ou virtuelle – de l'hyperonomase et de la phagocytose nous paraissent plus précis et plus adéquats. Ils permettent de rendre compte de l'assise de la traduction proposée en vertu du principe d'équivalence notionnelle défini au début de cet article. Ils sont certes plus difficiles à comprendre, mais leur rigueur nous paraît à la mesure des exigences du modèle notionnel.

4.4. Le R.N.I. face à l'approche viennoise

Le principe du R.N.I. trouve sa justification dans une approche terminologique qui prend en compte le terme, la notion et l'objet. Sous cet aspect, il demeure compatible avec le modèle triangulaire proposé par l'École de Vienne et contribue à insister sur le rôle prédominant des caractères dans la distinction des notions. Toutefois, la prise en compte des caractères dans le R.N.I. se double de l'observation des différences dans

25 Dans la tradition viennoise, la détermination, la conjonction et la disjonction sont les types de rapports de combinaison qui peuvent unir trois notions dans le cadre d'une relation logique TY (Felber, 1987 : 104-105). Notre modèle conduit à remettre en cause cette approche (Van Campenhoudt, 1994 : 103ssq.)

la manière dont chaque langue les appréhende, alors que dans le modèle viennois les caractères émanent d'objets matériels ou immatériels sur lesquels les langues n'ont censément aucune prise.

La perspective du R.N.I. est sous-tendue par une approche descriptive qui combine les démarches sémasiologique et onomasiologique. La première permet de dresser un inventaire des notions à partir des termes utilisés dans les diverses langues envisagées et de confronter les caractères activés. La seconde consiste à dénommer toutes les notions répertoriées dans le R.N.I., mais en préservant, si possible, l'intégrité référentielle dans chacune des langues. En effet, des mécanismes comme l'hyponomase, la phagocytose et la notion virtuelle permettent de respecter la manière dont la réalité est conçue et dénommée dans chaque langue.

4.5. Un modèle pour le terminographe ?

Il convient cependant d'observer que notre modèle est conçu à partir d'une terminographie particulière, orientée vers des objets essentiellement concrets. S'il s'applique fort bien à des réalités tangibles, aux frontières aisément identifiables, il ne pourrait prétendre rendre compte des équivalences entre les notions juridiques caractéristiques de la *Common Law* et celles propres au Code Napoléon.

Vue sous cet angle, cette modélisation doit avant tout être appréhendée comme un outil théorique permettant de mieux comprendre le mécanisme de l'équivalence et de l'analyser ponctuellement. Elle constitue également une piste de recherche pour l'élaboration d'un système informatique capable de mener rapidement un très grand nombre de raisonnements. Il est évident que la pratique de ce type de description n'est que d'un faible rendement pour le terminographe qui conçoit tout un dictionnaire et qui peut arriver au même résultat grâce à un travail soigné et respectueux du principe d'équivalence notionnelle.

Les mots-clés métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens

RUSSON WOOLDRIDGE et Isabelle LEROY-TURCAN

Université de Toronto, Canada et Université de Lyon III, France

Introduction

Les dictionnaires anciens mettent en œuvre une pluralité de systèmes de structuration textuelle, tant pour la macrostructure que pour la microstructure. Dans le domaine de la lexicographie française générale, le cas le plus marqué à cet égard est sans doute le *Thresor de la langue françoise* (= *TLF*) de Jean Nicot (1606), combinaison de dictionnaire monolingue, bilingue et multilingue, de dictionnaire de langue, dictionnaire encyclopédique et dictionnaire étymologique. Dans celui plus spécialisé de l'étymologie, le premier grand répertoire français, le *Dictionnaire étymologique, ou Origines de la langue françoise* (= *DEOLF*) de Gilles Ménage (1694), associe lui aussi les deux genres du dictionnaire étymologique et du dictionnaire général de langue au service d'une perspective ouverte, à visées linguistiques et encyclopédiques. Les articles individuels de ces deux ouvrages emploient différents modèles de contenu et d'articulation selon l'objet particulier de la description ou de l'analyse. Les modèles utilisés ne sont souvent qu'imparfaitement réalisés, ce qui crée un certain flou structurel, flou renforcé chez le lecteur par le caractère imprévisible des structures¹.

Aussi la base informatique qui a été réalisée pour le *TLF* de Nicot – celle du *DEOLF* de Ménage est en cours² – ne contient-elle, comme métatexte, que des jalons indiquant la localisation, les vedettes, la typographie, la langue des unités textuelles et les alinéas. Pour donner accès aux champs informationnels – catégorie grammaticale, définition, marque d'usage, exemple, citation, source, étymologie, etc. – sans dénaturer le texte original et sans empiéter sur les compétences de chaque lecteur en le

1. Sur la récurrence déficiente, cf. Wooldridge, 1977, sur la récurrence parfois prévisible grâce à une interprétation stylistique du texte dictionnaire, cf. Leroy-Turcan, 1994a et 1994b

2. Ce qui a été fait correspond à un corpus thématique (les végétaux) et à l'échantillon Ga. I, J, K, Ra-Re

conditionnant dans des pistes d'orientation ou dans des interprétations particulières, il a été élaboré des listes de mots-clés métalinguistiques, lesquels réunissent sous une forme lemmatique toutes les occurrences textuelles d'un marqueur de champ informationnel. Ainsi, par exemple, le lemme FEMININ permet de retrouver dans la base Nicot tous les contextes où le lexicographe a indiqué – par « féminin », « f. », « fem. », « foem. » ou « foemin. » – le genre d'un nom ou d'un adjectif féminin³.

Le projet d'informatisation des huit éditions complètes du *Dictionnaire de l'Académie française* (1694-1935) – projet annoncé à l'Institut de France en novembre 1994 (Wooldridge, 1994 ; cf. Leroy-Turcan et Wooldridge, 1995) – rencontre le même type de flou structurel que dans Nicot et Ménage, dans une mesure moindre mais bien réelle. Le texte du *Dictionnaire de l'Académie*, notamment celui de la première édition de 1694, quoique présentant une microstructure apparemment plus simple et plus récurrente que celles de Nicot et de Ménage, délimite souvent mal la frontière entre langue et métalangue, mot et référent, définition et marque d'emploi, locution, collocation et exemple. Ajouter un métatexte hautement structuré risquerait ainsi de dénaturer le texte en lui imposant une perspective moderne, donc anachronique⁴.

De fait s'opposent deux orientations radicalement différentes de balisage du texte informatisé selon les relations choisies : 1) via l'analyse du spécialiste qui propose une interprétation du fonctionnement du texte, ce qui se matérialise sous la forme d'un encodage complexe (cf. le balisage fin tel qu'il a été proposé pour Ménage et dont la mise en œuvre est complexe – Leroy-Turcan, 1994b) ; 2) via des grilles de lecture destinées à compléter le balisage minimal des marques de localisation et de typographie (cf. *supra* et *infra* les éléments retenus pour le balisage minimal) : les listes de mots-clés métalinguistiques.

Le but de notre communication est de donner une première mesure, fondée sur des échantillons informatisés⁵, de l'efficacité des mots-clés métalinguistiques, en relation avec la position microstructurelle et les marqueurs typographiques, comme outil d'interrogation structurante du *Dictionnaire de l'Académie* et, par extension, des dictionnaires anciens en général. La Base Académie Échantillon comprend, pour chaque édition, les tranches ÂME, DOUAIRE à DOUZIL, GAGNER, GRAS, GROS, LOIN à LOISIR, QUE, QUEUE, TIGE à TINTOUIN, VOLER. Un jalonnage indique l'édition, la vedette, l'alinéa, le caractère d'imprimerie et la page-colonne. À l'exception de la vedette, l'identification des champs balisés se fonde objectivement sur des critères formels systématiques. La communication démontre que le balisage explicite a été adopté comme méthode de recherche des vedettes plutôt que l'interrogation de la base à partir de marqueurs typographiques (capitales et position), alors que la localisation des champs informationnels se fonde sur le caractère d'imprimerie et les mots-clés métalinguistiques.

Un concept important pour cette méthode d'interrogation est celui de la « requête floue » (Wooldridge, 1993). En gros, le flou signifie que plutôt que de dépenser un effort énorme pour obtenir 100 % de ce qu'on cherche et rien de plus, on fait

3 Pour une discussion du concept de mot-clé métalinguistique, cf. Wooldridge, 1988 et à paraître

4 Cf. par exemple le cas des sous-vedettes non marquées dans Acad 1694 mais repérables en tant que telles grâce à Acad 1718.

5 Les différentes bases ont été créées et interrogées avec le logiciel WordCruncher

mieux et on obtient pratiquement les mêmes résultats en se contentant, avec beaucoup moins d'effort, d'une fourchette de 95 % à 105 % du total théorique, quitte à rejeter le 5 % de bruit. La requête floue convient particulièrement bien comme modèle d'interrogation d'un texte à flou structurel. Si nous introduisons la notion de mot-clé métalinguistique, c'est notamment en raison du flou macrostructurel des vedettes et sous-vedettes dans Acad 1694, flou qui touche aussi la microstructure.

1. Vedettes et sous-vedettes

Le système de classement des unités de la nomenclature employé dans la première édition (1694) est différent de celui des autres (1718-1935). Dans la première édition, les mots sont regroupés en familles étymologiques, lien caduque dans Acad2-8 ; dans la macrostructure principale, les mots de base des différentes familles sont rangés par ordre alphabétique, tandis que les autres membres de chaque famille sont organisés suivant des principes de dépendance dérivationnelle⁶.

Les propriétés formelles des unités de la macrostructure alphabétique sont les grandes capitales et la position initiale d'alinéa. À celles-ci s'ajoutent, dans certaines éditions, le gras (Acad6-8) et un interligne précédent plus grand (Acad8). Bien que les grandes capitales s'emploient aussi dans les renvois de la première édition (« TIMPAN. Voy TYMPAN. ») et que la position initiale d'alinéa puisse être occupée par des objets de toutes sortes, les deux ensemble suffisent en général à identifier toutes les vedettes et seulement des vedettes. Les rares exceptions sont à considérer comme des accidents⁷.

Le niveau secondaire de la macrostructure, celui des sous-vedettes, est plus problématique. Afin de rendre possible une comparaison du contenu de la macrostructure de la première édition avec les macrostructures des autres, il est nécessaire d'attribuer un jalon de vedette aux items subsidiaires d'Acad1 susceptibles d'appartenir à la nomenclature alphabétique d'Acad2-8 (ex. TIMIDITÉ, INTIMIDER, TIMIDEMENT ET TIMORÉ). Ces sous-vedettes ont deux propriétés formelles : petites capitales et position initiale d'alinéa ; cependant ces deux propriétés sont souvent partagées par ce qui dans l'ensemble des éditions (Acad1-8), comme dans la tradition lexicographique générale, doit être considéré comme des sous-adresses fonctionnant au niveau de la microstructure. Les difficultés posées dans Acad1 par la distinction des sous-vedettes et des sous-adresses sont nombreuses⁸.

6 Ainsi, pour les mots en TIM. la nomenclature principale donne TIMBALE, TIMBRE, TIMIDE, TIMON ET TIMPAN, alors que sous TIMIDE on trouve TIMIDE, TIMIDITÉ, INTIMIDER, TIMIDEMENT et TIMORÉ. (Est laissée de côté dans cette courte communication la question mineure des formes participiales, par ex. INTIMIDÉ) À partir de la 2^e édition, tous les mots sont donnés dans une seule macrostructure strictement alphabétique, ce qui a pour conséquence de séparer INTIMIDER, TIMIDE-TIMIDEMENT-TIMIDITÉ et TIMORÉ

7 Ex. QUI VIVE, dans l'alinéa « QUI VIVE. Voy VIVRE. » placé dans l'article qui, fonctionne de la même façon que QUICONQUE ou QUIDAM, comme sous-vedette de QUI.

8. Nous nous bornerons ici à donner une idée du problème en examinant les quelques premiers débuts d'alinéa en capitales de l'article LONG « LONG, LONGUE [...] SE FORLONGER [...] LOIN. [] AU LOIN [...] LOIN À LOIN, DE LOIN À LOIN. [...] LOINTAIN, AINE. [...] ÉLOIGNER. [...] ». Après application de la règle que la fin de la première unité d'un début d'alinéa en capitales est marquée par la première virgule à moins que celle-ci soit précédée d'un point, il nous reste : LONG, SE FORLONGER, LOIN, AU LOIN, LOIN À LOIN, LOINTAIN et ÉLOIGNER. Une comparaison avec Acad2-8 et la tradition lexicographique générale nous enseigne que LONG, FORLONGER (SE), LOIN, LOINTAIN et ÉLOIGNER sont des (sous-) vedettes et que AU LOIN ET LOIN À LOIN sont des sous-adresses de LOIN.

Comme les critères formels objectifs sont insuffisants pour permettre un jalonnage automatique des sous-vedettes d'Acad1⁹, une procédure raisonnable consiste à ajouter systématiquement un jalon de vedette à l'endroit de la première unité de chaque début d'alinéa en capitales, puis, dans une post-édition manuelle interprétative, à éliminer les jalons qui correspondent à une sous-adresse.

2. Caractères d'imprimerie et champs informationnels

Les deux principaux caractères d'imprimerie utilisés dans le *Dictionnaire de l'Académie* sont le romain et l'italique. Ils ont chacun des fonctions sémiotiques différentes, tout comme les capitales et les minuscules. Le gras ajouté aux vedettes d'Acad6-8 augmente la consultabilité du texte mais il est sémiotiquement redondant. Dans le système sémiotique de base, le romain minuscule (caractère non marqué) sert au niveau textuel fondamental du discours métalinguistique du lexicographe, lequel contient catégorie grammaticale, marque d'usage, filiation sémantique, définition et les copules articulatrices des différentes unités linguistiques et métalinguistiques de la microstructure ; les capitales romaines, l'italique et le gras (caractères marqués) s'emploient pour les autonymes – c.-à-d. les unités de l'objet de description, la langue : mots, expressions idiomatiques, cooccurrents, exemples, synonymes, etc. Mais peut-on se servir du caractère d'imprimerie, en rapport avec la position (position absolue ou relative d'un item dans la microstructure), pour rechercher les champs informationnels ?

Comme on l'a vu, le romain minuscule est utilisé pour plusieurs champs informationnels : la catégorie grammaticale est normalement signalée immédiatement après la vedette (« DOUBLE. adj. de tout genre. »), exceptionnellement ailleurs (« Il est aussi subst. ») ; les marques d'usage et de qualification sémantique tendent à être non initiales dans les alinéas discursifs des premières éditions (« On dit fig. et fam. [...] *une cervelle, une teste bien timbrée, mal timbrée* » Acad2-5 s.v. TIMBRER), initiales dans les dernières (« Fig. et fam., *Une cervelle, une tête timbrée* » Acad6-8). L'italique s'emploie systématiquement dans les exemples, de façon occasionnelle à l'endroit des cooccurrents et des synonymes : « GAGNER, se joint quelquefois avec la préposition *Sur* » (Acad2 ; Acad1 « *sur* ») vs. « SANS DOUTE, [...] se joint quelquefois avec *Que* » (*id.* ; Acad1 « *que* ») ; « DOUBLON, [...] On dit aussi, *Pistole* » (Acad6 ; Acad5 « [...] que nous appelons *Pistole* ») vs. « *Ne... que* peut, dans certains cas, être considéré comme entièrement synonyme de l'adverbe *Seulement* » (*id.* s.v. QUE = Acad7-8).

Le gras seul suffit pour trouver toutes les vedettes et co-vedettes d'Acad6-8 (325 séquences dans la Base Échantillon = 100 % des (co-)vedettes). Les grandes capitales romaines (387) sont utilisées pour des vedettes (Acad1-5) et des co-vedettes (Acad2-5) dans 374 cas (96,64 %) et pour des renvois dans 13 cas (Acad1, 3,36 %). Les petites capitales romaines (790) sont hautement polysémiques : elles servent régulièrement pour les co-vedettes d'Acad1 (5 occurrences = 0,63 %), les sous-vedettes et les sous-adresses d'Acad1-8 (726 = 91,90 %) et les renvois d'Acad4-8 (55 = 6,96 %) ;

⁹ Ce que confirme le cas de certains mots en italique qui fonctionneront ensuite comme adresse (cf. pour l'article LOUP, Leroy-Turcan et Wooldridge, 1995)

leur statut de caractère marqué expliquerait quatre occurrences idiosyncratiques, ou irrégulières : un synonyme (« On dit aussi *DUPLICATA* » Acad8 s.v. *DOUBLE*), un co-occurent (« Il se joint quelquefois avec la préposition *SUR* » Acad8 s.v. *GAGNER*) et un élément d'exemple (« *âme rachetée par le sang de JÉSUS-CHRIST* » Acad6-7 s.v. *ÂME* ; cf. Acad5 « [...] *JÉSUS-CHRIST* », Acad2-4 « [...] *Jésus-Christ* »).

Selon la logique générale du dictionnaire, les définitions (métalangue) sont imprimées en romain, les cooccurrents, synonymes et antonymes (langue) en italique. Lorsque, comme c'est souvent le cas des adjectifs et des adverbes, la définition est un mot plutôt qu'une périphrase, la distinction entre définition et synonyme est gommée ; en conséquence, l'emploi des différents caractères peut devenir aléatoire : « Il signifie aussi, *Espais*, et est opposé à *delié*, *delicat* » (Acad1-7 s.v. *GROS*), au lieu de « Il signifie aussi, *Espais*, et est opposé à *delié*, *delicat* » (cf. « *DOUBLE* [...] Il est opposé à *simple* » Acad1).

Il devient nécessaire alors d'avoir recours aux mots-clés métalinguistiques, tels que *SIGNIFIE*, *SE JOINT AVEC*, *ON DIT AUSSI*, *OPPOSÉ À*, etc., pour la recherche des définitions, cooccurrents, synonymes et antonymes.

3. Mots-clés métalinguistiques dans académie

La relative régularité, à travers les huit éditions, de l'emploi du caractère d'imprimerie s'observe aussi dans la terminologie du métalangage dictionnaire. Les noms sont normalement donnés comme noms masculins ou féminins, les verbes comme verbes transitifs ou intransitifs. Les formules de présentation des expressions lexicalisées, des définitions, de l'articulation sémantique et des niveaux d'usage restent les mêmes. En l'absence de l'étymologie et de la prononciation, qui n'est donnée que dans des cas exceptionnels, le nombre des champs informationnels est relativement petit. Pour la recherche des informations, quelques termes métalinguistiques sont caractéristiques par leur efficacité.

La Liste de mots-clés est un index alphabétique qui contient les adresses dans la base des occurrences des mots-clés métalinguistiques. Les items de la Liste sont des lemmes regroupant des formes variantes textuelles ; par exemple, le lemme *FEMININ* donne accès aux formes textuelles « f. », « fem. », « fém. », « féminin. » et « féminin ». La fréquence dans la Base Échantillon du mot-clé brut *FEMININ* est 204. 201 (98,53 %) des occurrences indiquent le genre de l'unité lexicale sujet d'énoncé ; dans presque tous les cas, le mot est précédé soit du mot-clé *SUBSTANTIF* (196), soit du mot-clé *ADJECTIF* (2). Un examen des six autres cas (6 sur les 8 occurrences de la forme « féminin ») révèle que trois d'entre eux concernent des signes linguistiques antonymes (« *GROSSE*, au féminin » Acad6-7 ; « Au féminin » Acad8 s.v. *GROS*), tandis que les trois autres se réfèrent à une propriété sémantique au niveau de la méta-métalangue (« On appelle en termes de Grammaire, *Noms douteux*. Ceux que les uns mettent au masculin, et d'autres au féminin. » Acad5-7). Le mot-clé *FEMININ* qualifiant une unité lexicale peut alors être corrigé pour en réduire le nombre d'occurrences à 201. Il est clair cependant que le mot-clé explicite *FEMININ*, tel qu'il vient d'être défini, ne donne pas accès à toutes les formes féminines de la nomenclature : le féminin de l'adjectif est normalement signalé par la forme elle-même et non pas par une étiquette (« *DOUX*, *DOUCE*. adj. »), tandis que les noms à genre double

sont marqués négativement par une absence d'étiquette de genre. Par exemple, les 150 occurrences du mot-clé SUBSTANTIF suivi ni de MASCULIN, ni de FEMININ renferment 30 concernant le féminin :

« TIGRE, TIGRESSE. s. » (Acad1-8)

« DOUILLET, est aussi substantif, dans la seconde acception. *Faire le douillet. C'est un douillet, une douillette.* » (Acad6-8)

Les occurrences du mot-clé FEMININ peuvent alors être augmentées par l'ajout des adresses des formes féminines de la nomenclature non étiquetées.

Dans le cas du genre, le manque d'une étiquette explicite n'exclut pas, comme nous venons de le voir, la recherche objective des items pertinents : TIGRESSE est donné comme féminin en vertu et de sa position comme seconde de deux co-vedettes et de l'indication « s. » ; DOUILLETTE est donné comme nom féminin dans l'exemple « *C'est une douillette.* ». Pour ce qui est du niveau d'usage et de la filiation sémantique, on doit se fier aux étiquettes explicites, sans lesquelles on est amené à faire une interprétation subjective du texte.

On peut adopter une méthode légèrement différente pour des termes comme « aussi ». Dans presque toutes ses occurrences en romain (770 sur 786 = 97,96 %), « aussi » est métalinguistique ; cette copule polysémique s'emploie dans des informations sur la catégorie grammaticale, le sens, la synonymie et la syntaxe. Pour la définition du mot-clé AUSSI, on a alors le choix entre la règle simple, floue mais efficace « 'aussi' précédé d'un jalon de caractère romain » (f 786), la liste globale plus précise des occurrences métalinguistiques (f 770) et la création de plusieurs mots-clés AUSSI correspondant à chaque type d'information particulier (catégorie grammaticale, sens, etc.).

L'usage familier est marqué comme tel dans le texte au moyen des termes « familial », « fam. », « famil. », « familière », « familières », « familiers » ou « familièrement ». Le mot-clé FAMILIER renvoie, dans la Base Échantillon, aux 319 occurrences de ces variantes. Il est important de distinguer la subjectivité de la décision du lexicographe de qualifier un item de familier – plutôt que, par exemple, de populaire¹⁰ ou de bas¹¹ – de l'objectivité de la recherche des étiquettes textuelles.

Pour rechercher les occurrences d'usage figuré, on peut choisir soit de se limiter au mot-clé FIGURÉ (formes textuelles « fig. », « figur. », « figuré », « figurées », « figurém. », « figurément » – f 517), soit d'y inclure ANALOGIE (« par analogie », « par une sorte d'analogie » – f 13) et/ou PROVERBIAL (« prov. », « proverb. », « proverbe », « proverbiale », « proverbialem. », « proverbialement » – f 275). On peut remarquer que dans 112 de ses occurrences PROVERBIAL se combine avec FAMILIER (ex. « On dit prov. et fig. *Jouer à quitte ou à double*, pour dire, *Hazarder tout pour se tirer d'une affaire.* » Acad1-5 s.v. DOUBLE ; cf. « [...] figurément et familièrement [...] » Acad6-7, « Voyez QUITTE » Acad8).

10. Le mot-clé POPULAIRE – « pop. », « popul. », « populaire », « populairement » – a une fréquence de 41.

11. Le mot-clé BAS – « bas », « bass. », « bassement » – a une fréquence de 19.

Si l'identification de la catégorie grammaticale, du genre et des marques d'usage est facile, celle d'autres champs informationnels, tels que la définition et l'exemple d'emploi, peut être complexe et nécessiter une interprétation subjective. Comme nous l'avons vu dans la section précédente, une condition préalable pour la définition est qu'elle soit en romain, pour l'exemple qu'il soit en italique. L'emploi de mots-clés métalinguistiques à l'endroit de ces deux types d'informations (SIGNIFIE, SE PREND POUR, COMME...) est occasionnel ; aucune combinaison de caractère d'imprimerie et de mots-clés ne permet de rechercher tous les cas de définitions/exemples et uniquement les définitions/exemples. Une considération préalable à l'application de marques formelles (ou à celle de jalons dans un dictionnaire moderne dont les champs informationnels ont été systématiquement balisés) est la définition de ce qui constitue une définition ou un exemple.

La « définition » peut fonctionner en métalangue de contenu ou en métalangue de signe (Rey-Debove, 1971) ; elle peut traiter le mot au niveau du lexique ou du discours :

Métalangue de contenu : « AME. s. f. Ce qui est le principe de la vie dans les choses vivantes. » (Acad1)

Métalangue de signe : « DOUBLE [...] se dit aussi des choses plus fortes, de plus grande vertu que les autres de mesme nature. » (*id.*)

Lexique : « ÂME [...] se dit aussi figurément de Ce qui est le principal fondement d'une chose, qui la maintient. *La discipline militaire est l'âme d'une armée. La bonne foi est l'âme du commerce.* » (Acad8)

Discours : « Fig., *Donner de l'âme à un ouvrage, mettre de l'âme dans un ouvrage*, Exprimer vivement ce qu'on y représente, y mettre beaucoup de feu, de sentiment. » (*id.* s.v. ÂME)

Les copules explicites reliant occasionnellement l'unité lexicale (sujet) à la définition (prédicat) comprennent « signifie » (f 414)¹², « pour dire » (f 801) et « se prend pour » (f 53).

« QUEUE, Signifie aussi, La dernière partie, les derniers rangs de quelque Corps, de quelque Compagnie » (Acad3)

« On dit, *Manger gras, faire gras*, pour dire, Manger de la viande les jours que l'on devroit manger maigre. » (Acad4 s.v. GRAS)

« Il se prend plus particulièrement, et d'une manière absoluë, pour *Façon d'agir douce*, et éloignée de toute sorte de violence. » (Acad1 s.v. DOUCEUR)

Une autre marque occasionnelle de la définition est l'explicitation du statut d'espece (hyponyme) de l'unité lexicale par opposition au genre (hyperonyme) du terme nucléaire de la définition. Ainsi, « espece/espèce » (f 78) et « sorte » (f 86) qualifiant, par exemple, DOUBLON et LOIR de types de monnaie et d'animal respectivement :

« DOUBLON. s. m. Espece de monnoye d'Espagne, qui est d'or, et que nous appellons *Pistole*. » (Acad1 et cf. Acad2-5) ; cf. « DOUBLON. s. m. Monnaie d'or espagnole qui a différentes valeurs. » (Acad6 et cf. Acad7-8)

« LOIR. s. m. Sorte de petit animal semblable à un Rat qui vit dans le creux des arbres et qui dort durant tout l'hyver, à ce que disent les Naturalistes. »

12. Cf. « sign. » (= « signifie » 9, « signification » 1) 10, « pour/peut signifier » 17, « signifiot/signifiat » 5, « phrases [...] qui signifient » 1, « signification(s) » 29.

(Acad1 et cf. Acad2-4) ; cf. « LOIR. s. m. Petit animal semblable à un rat, qui vit dans les creux des arbres, et qui dort durant tout l'hiver. » (Acad5 et cf. Acad6-8)

Pour ce qui est des exemples, il n'y a aucun moyen absolu de déterminer, dans le texte du dictionnaire, la frontière entre unités lexicales et exemples, entre syntagmes lexicalisés et syntagmes libres. Dans un alinéa qui contient plusieurs séquences en italique, les items lexicalisés précèdent normalement les items libres. Dans la plupart des cas, un syntagme lexicalisé est suivi d'un traitement sémantique, alors qu'un exemple libre est donné en position finale. Dans le premier extrait suivant, la première séquence en italique est une unité lexicale suivie d'une définition, la seconde une série de trois exemples ; dans le deuxième extrait, l'unique séquence en italique est une unité lexicale suivie d'une définition ; dans le troisième extrait, les multiples séquences en italique sont à considérer comme collocations ou phrases exemplificatrices même si plusieurs d'entre elles sont suivies d'une définition du mot en usage.

« On dit, *Filer doux*, pour dire, Demeurer dans la retenüe, dans la soumission à l'égard de quelqu'un que l'on craint, souffrir patiemment une injure. *C'est un homme avec qui il faut filer doux. je le feray bien filer doux. quand il s'entendit menacer, il fila doux.* » (Acad1 s.v. DOUX)

« On dit prov. *Aller doucement en besogne*. Et tantost il signifie, Sagement, meurement, sans rien précipiter ; tantost il signifie, Laschement, mollement. » (*id.* s.v. DOUCEMENT)

« DOUCEMENT. adv. d'Une maniere douce. *Dormir doucement. il faut marcher doucement dans la chambre d'un malade. heurtez doucement à la porte, c'est à dire avec le moins de bruit que l'on peut. Allez-y plus doucement. il faut traiter doucement les vaincus. reprendre quelqu'un doucement de ses fautes. je luy fis doucement la guerre de ce que, etc. quand on a souffert de grandes douleurs, et que l'on ne souffre plus, on se trouve bien doucement. on peut vivre doucement la campagne pour peu de chose. ce cheval galoppe fort doucement. cette affaire veut estre traitée, veut estre maniée doucement, c'est à dire delicatement. Il faut s'y prendre doucement. on craignoit qu'il n'arrivast quelque desordre dans l'assemblée : mais toutes choses s'y passerent fort doucement, c'est à dire fort paisiblement. C'est une chose qu'il faut faire doucement ; c'est à dire, sourdement, sans faire esclat.* » (*ibid.*)

Mais ce qui est valable pour l'Académie ne l'est pas forcément pour d'autres dictionnaires : ainsi les mots-clés métalinguistiques FEMININ et FAMILIER ne sont pas opératoires dans le cas du dictionnaire de Ménage qui n'aborde que rarement la synchronie.

4. Les variations fonctionnelles des mots métalinguistiques selon les genres de dictionnaires

Un même mot métalinguistique peut fonctionner différemment dans un dictionnaire de synchronie et dans un dictionnaire historique à dominante étymologique. C'est, par exemple, le cas de FEMININ dans Académie opposée à Nicot et à Ménage. Nous exa-

minerons, pour la marque de l'usage ancien, celui du mot métalinguistique ANCIEN.

Sous le lemme ANCIEN sont regroupées toutes les modalités de marques d'une graphie, d'un mot, d'une collocation ou d'un emploi qualifiés d'anciens (éventuellement par rapport à un usage en cours) ; sont donc compris sous ce lemme toutes les formes se rattachant à la base *ancien-* et les termes exprimant le même sémantisme comme *vieux* et *vieillir*, et leurs formes fléchies, ou les adverbes *autrefois*, *jadis*, sans négliger toutes les marques temporelles de passé dans les verbes qui peuvent être eux-mêmes métalinguistiques (comme *signifier*, *appeler* ou *dire*) ou éléments de définition (comme *valoir* s.v. DOUBLE : « Espece de monnoye qui valoit deux deniers » Acad2-5).

Le résultat des interrogations des trois bases – c.-à-d. le texte intégral de Nicot et des échantillons de Ménage et d'Académie (cf. *supra*) – donné sous forme de tableau (ci-dessous) nécessite quelques commentaires en raison des difficultés d'appréciation liées à la nature même de chaque dictionnaire.

	Nicot 1606	Ménage 1694	Acad 1694-1935
<i>ancien-</i>	234	24	15
<i>vieux</i>	3	11	11
[il] <i>vieillit / a vieilli</i>	0	0	14
<i>autrefois</i>	0	4	15
<i>jadis</i>	19	1	0

Principales marques d'usage ancien

Ancien- dans Ménage. Sur 90 occurrences d'*ancien-*, 49 ne sont pas du tout pertinentes, 12 concernant un discours socio-culturel, 7 étant dans des citations et 30 appartenant à la bibliographie ; les occurrences restantes se répartissent entre l'étymologie (sur 10 occurrences, 2 étymons = « ancien mot » ; 4 renvois à d'autres langues dont 3 séquences « de l'ancien » ; un emploi « d'ancienne origine »), des références à l'ancien français (3 occurrences = « mot ancien ») et l'usage ancien (10 emplois d'*anciennement* tous combinés à des marques d'imparfait et 14 d'*ancien*), sans compter les doubles emplois dans un même article. La diversité des occurrences d'*ancien-* rend nécessaire la définition des différentes conditions de l'environnement du mot métalinguistique réparti dans des sous-catégories de séquences métalinguistiques levant toute ambiguïté.

Vieux dans Ménage. Sur les 21 occurrences de *vieux* dans Ménage, seulement 11 sont pertinentes pour l'identification d'un usage ancien ; 4 emplois qualifient des références bibliographiques, un emploi qualifie un nom de poète, un autre un proverbe, 3 se trouvent dans des citations, 5 emplois de la forme du féminin n'appartiennent pas à la métalangue, un emploi concerne l'étymologie ; la proportion importante de rebut nous conduit à proposer des modalités de structuration ou de modélisation de l'environnement du mot métalinguistique susceptible, dans ce cas, de devenir plutôt une séquence métalinguistique qui inclut les éléments textuels permettant une interrogation plus fine. De fait, l'interrogation par la séquence « Vieux mot », en début d'article, ou « , vieux mot », en groupe apposé, donne le résultat des 11 occurrences pertinentes, s.v. GABAN, GABER, GALLER, GAUSSER, JOUCARITE, JUS, ISNEL, RAIN, RAMON, RAMPONNER ET RESE.

Des problèmes analogues s'observent dans le discours fortement étymologique de Nicot, la proportion des remarques d'usage restant dominante (les 234 occurrences d'*ancien-* sont à trier).

On peut faire le même genre d'analyse pour les séquences « on dit », « on disoit », « on a dit », qui n'ont pas le même fonctionnement dans Nicot, Ménage et Académie.

Conclusion

Bien que les dictionnaires modernes ne soient jamais entièrement systématiques, ils le sont relativement ; quand on les informatise par *rétroconversion*, on balise systématiquement leurs champs informationnels à un degré plus ou moins détaillé. Les dictionnaires anciens sont, dans une mesure variable, moins systématiques que les dictionnaires modernes. Pour ne pas les enfermer dans une interprétation univoque, on doit éviter un balisage systématique des champs informationnels. En revanche, on peut, dans la majorité des cas, obtenir un taux de succès très satisfaisant dans la recherche des champs informationnels au moyen des indicateurs que sont le caractère d'imprimerie et les mots-clés métalinguistiques. Dans l'utilisation des jalons de caractère et la définition des mots-clés, il faut réfléchir au rendement de la recherche floue par opposition à une post-édition ardue : la seule interrogation de ce genre de base par les mots métalinguistiques ne saurait produire des statistiques utilisables de façon automatique ou manuelle pour la fréquence ou le repérage des champs informationnels ; même une définition rigoureuse des différentes modalités d'environnement du mot métalinguistique exige les compétences linguistique, dictionnaire et pragmatique du lecteur/utilisateur de la base.

Représentation de la polysémie dans un dictionnaire électronique

Michel MATHIEU-COLAS

Laboratoire de Linguistique Informatique, Université Paris XIII - CNRS - INaLF, Villetaneuse, France

• Abstract •

The traditional treatment of polysemy causes many problems of representation : the coexistence of several meanings within the same entry generates extremely complex structures, which are difficult for a computer to exploit. In developing electronic dictionaries, our suggestion would be to generalize and to systematize the splitting up (dégrouper) of the entries : it is better for each use to be considered as a full « word », i.e. to receive an independent address and a specific description (morphological, syntactic and semantic information, translations, etc.). The model allows as many descriptions as there are meanings. This does not necessarily lead to dispersion of information : relations between the different uses (branchings, shifts in meaning ..) can be reintroduced into appropriate fields, which allows a more flexible representation of polysemic connections.

Étant admis que la polysémie constitue une donnée fondamentale des langues naturelles et l'une des principales difficultés pour le traitement automatique, nous nous interrogerons ici plus particulièrement sur les modalités de *représentation* de cette pluralité dans le cadre des dictionnaires électroniques. Après un bref rappel de la conception lexicographique classique, nous plaiderons en faveur d'une nouvelle disposition des entrées de dictionnaire et tenterons de répondre aux objections que pourrait soulever le modèle proposé.

Précisons que ces réflexions s'inscrivent dans le cadre des recherches que nous menons, avec Gaston Gross, au Laboratoire de Linguistique Informatique de Villetaneuse (LLI). Si certaines de nos propositions peuvent paraître évidentes en terminologie, elles le sont peut-être moins du point de vue linguistique et lexicographique, à en juger par la diversité des approches¹.

1. Pour une autre approche de la polysémie, voir ici même la communication de Pierrette Bouillon.

1. La conception classique

Quelques remarques suffiront à illustrer le traitement traditionnel de la polysémie. La lexicographie classique repose sur une conception fondamentalement **unitaire** du mot : tous les emplois sont regroupés au sein d'un même article, la multiplicité étant prise en charge par différents systèmes de hiérarchisation et de classement des sens. Seuls sont exclus de ce système les véritables homonymes, fondés sur des étymons différents (les trois mots *baie* qui coexistent en français), et quelques familles anciennement éclatées, à l'instar de *voler* (*to fly / to steal*) ou de *grève*.

Il en résulte, pour les termes polysémiques, de nombreux problèmes de représentation : disposition linéaire ou arborescente, ordre logique ou historique, opposition entre langue générale et langues de spécialité, etc. Quels que soient les choix effectués, la coexistence de plusieurs emplois au sein d'un même article se traduit par des structures d'une grande complexité et au surplus très différentes d'un dictionnaire à l'autre.

Il est vrai que ce modèle unitaire a connu, récemment, quelques aménagements. L'exemple le plus familier en est sans doute le *Dictionnaire du français contemporain* (DFC, Larousse, 1971), où de nombreux termes polysémiques sont dégroupés, ce qui revient à les traiter comme de simples homographes : des lexèmes comme *bureau*, *cher* ou *commander* se trouvent ainsi décomposés en deux ou trois entrées, ce qui a pour effet de faciliter la description synchronique des emplois et de permettre un traitement plus rigoureux des séries dérivationnelles.

Mais il ne s'agit là que d'une solution de compromis, car l'on s'arrête à mi-chemin : au sein de chaque entrée peut subsister une multiplicité d'emplois qui reproduit, au second degré, le modèle traditionnel. Si *bureau*, désormais, a droit à deux adresses, chacune d'elles n'en est pas moins subdivisée en trois ou quatre descriptions :

1. **bureau** n.m 1^o Table, munie ou non de tiroirs, dont on se sert pour écrire [...] – 2^o Pièce où est installée cette table [...] – 3^o Mobilier de cette pièce [...]
2. **bureau** n.m. 1^o Établissement public où sont installés des services administratifs [...] – 2^o Caisse d'un théâtre [...] – 3^o Ensemble des employés ou des fonctionnaires qui travaillent dans une administration [...] – 4^o Membres d'une assemblée, d'une association, élus pour diriger les travaux [...]

Les difficultés liées à la polysémie demeurent ici entières : le problème est déplacé, il n'est pas résolu.

2. Les mérites du dégroupement

Nous proposons en conséquence de systématiser et de généraliser le principe du **dégroupement** : même lorsqu'il s'agit de langue générale, chaque emploi gagne à être considéré comme un « mot » à part entière, ce qui revient à lui attribuer une adresse autonome et une description spécifique (*bureau* donnerait ainsi lieu, pour reprendre l'exemple précédent, à sept entrées distinctes). Pratiquant cette technique depuis quelques années, dans le cadre des travaux du LLI, nous sommes de plus en plus conscients des avantages qu'elle offre.

Rappelons que notre conception des dictionnaires électroniques s'inspire largement de la pratique des bases de données : chaque entrée constitue un « enregistrement », cependant que la description se trouve répartie en une série de « champs » (rubriques) clairement définis, correspondant aux différents paramètres de l'information lexicographique. Voici, à titre d'exemple, l'ébauche de description d'une unité monosémique (*taille-crayon*) :

MOT :	taille-crayon
CAT. GRAM. :	nm
STRUCTURE :	v00 [<i>verbe + nom</i>]
FLEXION :	00 ; 01
VARIANTES :	taille-crayons
TRAITS :	inc [<i>inanimé concret</i>]
CLASSE :	instrument
DOMAINE :	écrit., dess.
ANGLAIS :	pencil sharpener
ALLEMAND :	Bleistiftspitzer

On trouve ici représentées des données morphologiques (structure formelle, flexions, variantes graphiques), des informations sémantiques (domaines) et/ou syntaxiques (les traits et, plus précisément, ce que Gaston Gross et moi appelons les « classes d'objets² »), ainsi que des traductions. Les mêmes informations peuvent être visualisées dans d'autres formats structurellement équivalents, notamment sous forme linéaire, les rubriques étant délimitées par des séparateurs et des identificateurs :

taille-crayon /G:nm /M:v00 /F:00; 01 /V:taille-crayons /T:inc /C:instr. /D:écrit.,dess. /AN:...

ou sous forme de tableau, les lignes et les colonnes correspondant respectivement aux entrées lexicales et aux champs de description :

MOT	G:	M:	F:	V:	T:	C:	D:
TAILLE-CRAYON	nm	v00	00.01	taille-crayons	inc	instr.	écrit .dess
TAILLEUR-PANTALON	nm	nm00	01-01		inc	vêt fém	habill
TALK-SHOW	nm	d62	01		évé.	émission	télév
TALKIE-WALKIE	nm	d62	01-01		inc	appar.	radiocomm

2. G. Gross, 1994, M. Mathieu-Colas, 1994, pp 162-173

Ces exemples simplifiés ne rendent pas compte, naturellement, du nombre réel des paramètres qui articulent nos analyses. Ainsi, pour les mots prédicatifs, nous indiquons la structure argumentale (y compris les traits et les classes qui spécifient les arguments : voir *infra*), à quoi s'ajoute, pour les noms abstraits, l'indication des verbes supports (*voyage* se construit avec « faire », *ordre* est introduit par « donner »). S'agissant des informations sémantiques, nous mentionnons, quand il y a lieu, les synonymes, les antonymes, les relations méronymiques (relations partie-tout : Otman, 1995 : chap. 6), de même que nous notons, sur le plan pragmatique, les registres temporels, régionaux ou sociaux. La liste n'est pas close, et d'autres informations pourraient ici trouver leur place : indication des dérivés, « fonctions lexicales » (I. Mel'cuk), indices de fréquence, etc.

Quel que puisse être dans le détail le choix des rubriques, on retiendra surtout, pour la présente analyse, l'importance de la *fiche* en tant que principe organisateur de l'information lexicographique (ce qui nous rapproche de son utilisation en terminologie : voir Lerat, 1990). Ce mode de structuration comporte deux aspects complémentaires : d'une part, il implique que les données lexicales puissent être décomposées en paramètres discrets formalisables (fondant ainsi la possibilité de procéder à des extractions et des traitements automatiques) ; d'autre part, il signifie que chaque entrée du dictionnaire correspond à un emploi strictement défini. Il en résulte qu'on est conduit, en cas de polysémie, à développer **autant de descriptions qu'il y a de sens différents** – soit par exemple, pour le mot *crapaud* :

MOT	TRAIT	CLASSE	DOMAINE	REGISTRE	ANGLAIS
CRAPAUD #1	ani	batracien	zool		<i>toad</i>
CRAPAUD #2	hum	qualif		fam. (gamin)	<i>brat</i>
CRAPAUD #3	hum	qualif		fam. (pers. laide)	
CRAPAUD #4	ina	mal. anim.	vétér.		<i>greasy heel</i>
CRAPAUD #5	inc	instr. mus.	mus.		<i>baby grand</i>
CRAPAUD #6	inc	siège	ameubl.		<i>tub easy-chair</i>
CRAPAUD #7	inc	dispos.	ch. de f.		<i>sleeper clip</i>
CRAPAUD #8	inc	dispos.	pyrotechn.		<i>jumping cracker</i>
CRAPAUD #9	inc	défaut	joaill.		<i>flaw</i>
CRAPAUD #10	inc	support	topogr.		
<i>etc.</i>					

Le dégroupement ainsi conçu a le mérite de la simplicité, tant du point de vue linguistique (clarté et lisibilité) que sur le plan informatique (facilité de traitement pour la machine). Plus particulièrement, chaque paramètre de la description est susceptible, par ce moyen, de gagner en précision.

a) Cela vaut déjà, d'une certaine manière, pour les informations morphologiques : chaque emploi est susceptible d'avoir son propre genre (*un espace/une espace*), son type de conjugaison (*saillait/saillissait*), ses variantes graphiques (*porte-aiguille[s]* en couture, mais non en chirurgie), sa mise au féminin :

« *VENDEUR, EUSE* n. 1. Personne dont la profession est de vendre, en partic. dans un magasin.
2. DR. Personne qui fait une acte de vente. (En ce sens, le fém. est *venderesse*.) »
(PETIT LAROUSSE)

Le souci d'unité conduit ici au paradoxe (la parenthèse finale contredit la formulation de l'entrée), alors que le dégroupement que nous proposons permet de décrire plus simplement chacun des deux emplois :

vendeur #1	/D:comm.	/F:6B (= fém. <i>vendeuse</i>)
vendeur #2	/D:dr.	/F:68 (= fém. <i>venderesse</i>)

Le même problème peut se poser pour la mise au pluriel, comme l'illustre l'entrée *œil, yeux* du TLF, contredite par une remarque livrée en fin d'article : « Dans les sens techn., le plur. de *œil* est *œils* : *les œils d'une voile*. » Mais rien ne permet de savoir, dans le détail, à quels emplois précis s'applique cette remarque. Le dégroupement, au contraire, rend à chacun son dû :

œil #1	/D:lg	/F:06 (= plur. <i>yeux</i>)
œil #2	/D:arm.	/F:01 (= plur. <i>œils</i>)
œil #3	/D:bourell.	/F:01
œil #4	/D:hort.	/F:06
œil #5	/D:impr.	/F:01
œil #6	/D:jeux (go)	/F:06

b) Les avantages du dégroupement sont plus sensibles encore, naturellement, pour la composante proprement sémantique de la description. La notation, pour chaque sens, de toutes les informations pertinentes (classes, domaines, marques d'usage, etc.) assure une représentation plus fine de la polysémie et facilite, en conséquence, les procédures de levée d'ambiguïtés.

Un seul exemple, situé aux confins de la sémantique et de la syntaxe, suffira à illustrer notre propos : il s'agit de la construction des termes « prédicatifs » (verbes, adjectifs, noms abstraits), que nous évoquions précédemment. Soit le verbe *conduire* et les deux phrases suivantes :

Ce sentier conduit à la mer
Pierre conduit un poids lourd

Chacun des deux emplois se caractérise par une distribution spécifique : sujet <voie de communication> et complément locatif pour le premier, sujet humain et objet <véhicule> pour le second. On remarquera en particulier l'intérêt des « classes d'objets » pour la définition des arguments (si l'on omet de préciser que le sujet de la première phrase est une « voie », on s'expose à générer des phrases non acceptables : **ce crayon/*ce bébé/*cette surprise conduit à la mer*). Or de telles informations sont plus aisées à représenter si chaque emploi bénéficie d'une description différenciée :

conduire #3	/N0: <voie>	/N1: loc	(<i>Ce sentier conduit à la mer</i>)
conduire #12	/N0: hum	/N1: <véhic>	(<i>Pierre conduit un poids lourd</i>)

Il en irait de même pour un adjectif comme *juste* :

juste #1	/N0: hum	(<i>Cet examinateur est juste</i>)
juste #5	/N0: <instrum. de mesure>	(<i>Cette balance est juste</i>)
juste #6	/N0: <instrum. de musique>	(<i>Ce piano n'est plus très juste</i>)
juste #8	/N0: <vêtement>	(<i>Cette veste est un peu juste</i>)

Le dégroupement, ici et là, est source de clarté.

c) Les mêmes principes inspirent notre traitement des mots composés et des locutions (plusieurs dizaines de milliers d'unités sont en voie de description). Dans les dictionnaires que nous élaborons, chaque unité complexe possède sa propre entrée et fait l'objet d'un traitement distinct (par exemple *carte bleue*, *carte orange*, *carte grise*, *carte verte*, *carte de séjour*, *carte de travail*, *carte d'électeur*, *carte de crédit*, *carte de visite*, *carte d'identité*, *carte à jouer*, *repas à la carte*, *jouer cartes sur table*, *donner carte blanche* à <Nhum>, etc.), ce qui permet d'élaborer des descriptions plus précises.

d) Enfin, *last but not least*, le traitement multilingue tire directement avantage du dégroupement : chaque emploi faisant l'objet d'une description spécifique, il suffit d'indiquer, pour chaque entrée, la ou les traductions appropriées (revoir *supra* l'exemple du mot *crapaud*).

3. Objections et réponses

Le dégroupement systématique auquel nous proposons de recourir est toutefois susceptible de soulever un certain nombre d'objections. Nous nous limiterons ici à deux critiques majeures.

3.1. Continuité ou discontinuité

La première concerne notre conception *discontinue* de la polysémie : nous traitons les mots comme des ensembles d'emplois discrets et clairement différenciés. Or on sait que d'autres travaux privilégient plutôt une représentation *continuiste* du sens (voir, par exemple, Kayser, 1987 et Fuchs, 1988). S'opposant au point de vue « homonymique » fondé sur une pluralité de significations disjointes, B. Levrat (1993) plaide pour une « optique polysémique » et postule une « signification unique », un « sémantisme de base » qui s'enrichit sous l'influence du contexte pour donner naissance à un ensemble de valeurs apparentées.

Cette question, trop fondamentale pour que nous puissions l'évoquer en quelques lignes, mériterait à elle seule de plus amples développements. Disons seulement, pour notre propos, que de telles recherches offrent un grand intérêt du point de vue de la constitution d'une théorie du sens, mais semblent très complexes à mettre en œuvre dans la pratique et difficiles à appliquer à une grande échelle (quand il s'agit de décrire plusieurs dizaines de milliers de mots). En outre, même du point de vue théorique, nous rejoindrions volontiers les analyses de R. Martin, pour qui les figures de *surdétermination*, d'*indétermination* et de *neutralisation* décrites par C. Fuchs « ne

peuvent se définir qu'à partir de sens ou d'acceptions préalablement distingués » (Martin, 1994 : 92). Même les métaphores ou les métonymies les plus audacieuses, les créations poétiques les plus libres ne contredisent pas l'existence d'emplois plus stables organisés de façon discrète ; au contraire, ils les présupposent, ils prennent appui sur eux pour mieux produire leurs effets par un jeu subtil d'allusions, de détournement et de reconstruction du sens. Ce sont les emplois stables et lexicalisés qui constituent, avant toute chose, l'objet des dictionnaires : dans cette perspective, le traitement lexicographique de la polysémie s'accommode mieux, nous semble-t-il, d'une représentation discontinue.

3.2. Articulation des emplois

La deuxième objection que nous voudrions examiner met en cause, par-delà la discontinuité, l'émiettement et l'éclatement des descriptions, et la perte d'information qui pourrait en résulter. Même si, dans un dictionnaire traditionnel, les divers sens d'un mot sont clairement différenciés (hypothèse discontinuiste), ils demeurent **articulés** les uns aux autres dans l'unité du mot (matérialisée par la cohésion de l'article) ; à l'inverse, dans notre système, le lien semble rompu : on perçoit bien ce qui distingue les emplois, on ne voit plus ce qui les relie... Cette objection, à notre sens, n'invalide pas le dégroupement, mais conduit au contraire à un enrichissement et à un approfondissement du modèle.

Qu'on nous permette de commencer par un exemple artificiel. Supposons qu'un mot se trouve ainsi décrit :

MOT I.	Sens A.
II.	<i>Par anal.</i> sens B.
III. 1.	<i>Par méton.</i> sens C.
2.	<i>Spécialt.</i> sens D.

L'application du dégroupement, tel que nous l'avons défini, conduit à un « aplatissement » de la description, donc à un appauvrissement :

mot #1 sens A
mot #2 sens B
mot #3 sens C
mot #4 sens D

Rien n'empêche cependant de réintroduire ici l'information relationnelle, de façon structurée et explicite, en consignand dans un champ supplémentaire les éventuelles dérivations sémantiques :

ENTRÉES	EMPLOIS	DÉRIVATION SÉMANTIQUE
<i>mot #1</i>	sens A	
<i>mot #2</i>	sens B	< 1 (analogie)
<i>mot #3</i>	sens C	< 1 (méton.)
<i>mot #4</i>	sens D	< 3 (spécial.)

Les indications fonctionnent ici comme un système de « pointeurs » permettant de relier, de proche en proche, les entrées associées : l'emploi n° 4 dérive, par spécialisation de sens, de l'emploi n° 3, lequel procède, par métonymie, de l'emploi n°1... On peut aussi, si on le souhaite, réintégrer dans la description l'arborescence sous-jacente au modèle classique – ces rapports hiérarchiques qui s'expriment traditionnellement sous forme de lettres ou de chiffres (I.A.1.a...) :

ENTRÉES	EMPLOIS	DÉRIVATION SÉMANTIQUE	ARBRE
<i>mot #1</i>	sens A		I
<i>mot #2</i>	sens B	< 1 (analogie)	II
<i>mot #3</i>	sens C	< 1 (méton.)	III.1
<i>mot #4</i>	sens D	< 3 (spécial.)	III.2

On voit ainsi qu'il est possible, dans le cadre d'une présentation dégroupée, de représenter toute l'information relative aux liens logiques ou historiques qui articulent les emplois.

Cela pourtant ne suffit point. Nous voudrions montrer que le format proposé ne permet pas seulement une simple « récupération » d'informations déjà présentes par ailleurs, mais qu'il apporte en outre une amélioration. Il n'est, pour s'en convaincre, que d'observer l'état présent de la lexicographie : d'un dictionnaire à l'autre, on note une assez grande stabilité dans les emplois décrits, mais une extrême disparité dans leur disposition (variations affectant l'ordre et la hiérarchie des acceptions). Voici comment trois dictionnaires – *Petit Robert*, *Petit Larousse illustré*, *Dictionnaire du français contemporain* – présentent le mot *colle* :

<i>PR</i>	<i>PLI</i>	<i>DFC</i>
1. Substance	1 Substance	1. Substance
2. <i>Arg scol</i> Interrogation	2 <i>Arg scol</i>	2.1. <i>Fam.</i> Question embarrassante
<i>Cour</i> Question difficile	a) Interrogation	2.2. <i>Arg. scol.</i> Interrogation
Punition	b) Punition	3. <i>Arg scol</i> Punition
	3. <i>Fam.</i> Question embarrassante	

Les mêmes emplois sont présentés, mais la structuration diffère. Au risque d'employer une image paradoxale, on pourrait dire qu'ici, les feuilles de l'arbre sont constantes : seule change l'arborescence.

Une telle observation est parfaitement compréhensible du point de vue linguistique : les emplois, en effet, reflètent directement la pratique de la langue, ils constituent, pourrait-on dire, sa réalité première – alors que l’organisation lexicographique, relevant du niveau métalinguistique, est une structure au second degré, sujette comme telle à interprétation et à révision. D’où l’intérêt qu’il peut y avoir à dissocier, dans la présentation, les deux niveaux de structuration.

Supposons que l’on veuille modifier, d’une édition à l’autre, la disposition des emplois, par exemple remplacer l’arborescence du *DFC* par celle du *Petit Larousse*. Cela supposerait, dans le format traditionnel, une refonte complète de l’article. Dans un dictionnaire électronique tel que nous le concevons, il suffit de modifier l’information relationnelle dans le champ approprié :

		<i>DFC</i>	<i>PLI</i>
<i>colle</i> #1	substance	1	1
<i>colle</i> #2	question embarrassante	2.1	3
<i>colle</i> #3	interrogation scolaire	2.2	2a
<i>colle</i> #4	punition	3	2b

Une telle souplesse de traitement est susceptible de faciliter la maintenance des dictionnaires.

On peut aussi, si on le souhaite, juxtaposer plusieurs dispositions dans des champs différents – l’un reflétant l’ordre historique, l’autre figurant l’arborescence logique, etc. – ce qui est évidemment impossible dans le schéma classique ; ou encore, dans le cadre d’une procédure de désambiguïsation automatique, attribuer aux emplois un ordre séquentiel correspondant à un algorithmique de décision (Martin³, 1994) ; ou même choisir de n’imposer aucune hiérarchie (comme le fait le *GDEL* pour le mot *colle*). Les emplois constituant la partie stable de l’édifice, les liens qui les unissent peuvent être définis (ajoutés, modifiés, supprimés) avec toute la liberté souhaitable.

Nous concluons par un point de méthode. Pour un phénomène aussi complexe que la polysémie, il s’avère plus opératoire, du point de vue linguistique comme du point de vue informatique, de commencer par rendre compte de la **diversité** des éléments (dégrouper maximal) avant de pouvoir décrire, avec plus de précision, les **liens** qui les unissent. Disons-le autrement : étant donné une structure à la fois une et multiple – comme l’est la polysémie –, il est techniquement plus simple de partir du multiple pour y introduire l’unité que d’effectuer l’opération inverse. Je ferais volontiers mienne la devise épistémologique du philosophe Jacques Maritain : *distinguer pour unir*⁴.

3. R. Martin, 1994, pp. 101 et suiv. Une lecture attentive de l’article *remettre* dans le *TLF* permet à l’auteur d’identifier une quarantaine d’emplois distincts et de les réorganiser systématiquement dans la perspective d’un traitement automatique (construction d’un algorithme permettant la sélection des sens par un automate).

4. Voir le volume paru en 1932 sous le titre *Distinguer pour unir ou Les degrés du savoir* (Bibliothèque française de philosophie, Paris, Desclée de Brouwer)

Une base de données lexicale multilingue interactive

Catherine WALTHER et Éric WEHRLI

Laboratoire d'analyse et de technologie du langage (LATL), Département de linguistique, Université de Genève, Suisse

1. Introduction

Les applications liées au traitement automatique du langage (TAL) exigent d'une manière générale de très gros lexiques, de l'ordre de plusieurs dizaines, voire de plusieurs centaines de milliers d'entrées. Or, l'établissement d'une base de données lexicale est une entreprise de longue haleine, exigeant des moyens importants. On comprend mieux, dès lors, que deux préoccupations majeures dans l'établissement des lexiques au cours de ces dernières années aient été, d'une part, l'utilisation de dictionnaires informatisés et, d'autre part, la réutilisabilité. La première de ces tendances vise à extraire des dictionnaires conventionnels disponibles sur support informatique les données propres à l'élaboration de lexiques utilisables pour le TAL (cf. Atkins et Zampoli (1994) ; Boguraev et Briscoe (1988)). Quant à la seconde tendance, elle tend à définir le contenu des lexiques de façon suffisamment générale (c'est-à-dire non spécifique à une application particulière) de façon à faciliter l'utilisation des lexiques pour d'autres applications. Idéalement, on souhaiterait disposer d'un lexique susceptible d'être utilisé, moyennant quelques inévitables ajouts, pour toute une série d'applications pour une langue ou un groupe de langues données.

C'est dans ce contexte, et avec ces objectifs que le LATL a entrepris, il y a quelques années, le développement d'une base de données lexicale multilingue. Limitée dans un premier temps au français et à l'anglais écrits, cette base de données a été augmentée à plusieurs reprises, d'une part par l'ajout de l'allemand, d'autre part par celui des données nécessaires au traitement de la langue orale. L'approche adoptée s'est voulue résolument pragmatique, et les ajouts ont été effectués en fonction des besoins des applications. Initialement développée pour servir des applications dans le domaine de l'analyse syntaxique et de la traduction interactive, cette base de données est maintenant utilisée pour des projets aussi divers que l'étiquetage (*tagging*), la synthèse vocale à partir de textes (*text-to-speech*), l'établissement de concordances lemmatisées et, dans un avenir proche, l'analyse de la parole.

Dans cet article, nous décrivons la structure de cette base de données lexicale

multilingue interactive (LMI), qui comprend à ce jour des dictionnaires monolingues pour le français, l'anglais et l'allemand, les dictionnaires bilingues (de transfert) correspondants, ainsi que des dictionnaires d'utilisateur et des dictionnaires spécialisés, construits sur la même architecture. Les dictionnaires monolingues et bilingues ont été adaptés de manière semi-automatique (interactive) à partir de plusieurs dictionnaires informatisés, dont l'*Oxford Advanced Learner Dictionary of Current English*, le *Micro Robert-Collins*, le *Grand Robert*, et les bases de données CELEX, Brulex et BDlex.

2. Structure des dictionnaires monolingues

Les dictionnaires monolingues français, anglais et allemand sont tous organisés sur le modèle de la spécification morphologique complète, ce qui signifie qu'il s'agit de dictionnaires de mots et non de morphèmes (cf. Jackendoff (1975) et Wehrli (1985)). Les diverses formes morphologiques associées à un lexème particulier correspondent à des entrées distinctes mais liées les unes aux autres, les relations morphologiques étant exprimées sous la forme d'un ensemble de relations entre entrées lexicales indépendantes. Ce mode d'organisation s'éloigne radicalement des modèles plus classiques utilisés en linguistique informatique, dans lesquels la morphologie constitue habituellement une composante indépendante entre le dictionnaire et l'analyseur ou le générateur.

Comme toutes les variantes morphologiques sont présentes en tant que telles dans les dictionnaires monolingues du système, il n'est pas nécessaire de décomposer les mots en analyse, ou de recomposer des formes dérivées en génération.

À l'usage, cette base de données s'est avérée particulièrement efficace (la recherche lexicale se réduit à une recherche dans la base de données), fiable (le système ne peut produire que des formes existantes) et souple (LMI peut être utilisée pour pratiquement n'importe quel type d'application). De plus, le fait que les dictionnaires monolingues contiennent toutes les formes fléchies permet d'associer à chaque variante morphologique d'un mot les informations qui lui sont propres, comme la représentation phonétique, la structure syllabique ou encore sa fréquence.

Il convient pourtant de noter que les dictionnaires monolingues ne se réduisent pas à une simple liste de mots dans laquelle les traits syntaxiques et sémantiques seraient répétés pour tous les membres d'un paradigme morphologique. Dans le but d'éviter ce genre de redondance, nous distinguons deux types d'entrées, soit les formes orthographiques de surface (ou *mots*), et des formes d'un niveau plus abstrait qui correspondent aux lectures des mots (ou *lexèmes*). Les expressions idiomatiques (ou *idiomes*) constituent le troisième type d'entrée des dictionnaires monolingues et bilingues dans la base de données LMI.

2.1. Information liée aux mots

Les entrées de type *mot* contiennent l'information typiquement associée aux formes orthographiques de surface, comme les traits grammaticaux et d'accord (catégorie lexicale, nombre, genre, cas, temps, etc.), la représentation phonétique du mot (y com-

pris la structure syllabique, les accents, pour le français la consonne latente à réaliser dans les contextes de liaison, etc.), et la fréquence d'occurrence. À titre d'exemple, les informations associées avec une forme verbale (*mangeront*) ou adjectivale (*grand*) sont illustrées partiellement en (1) et (2).

- (1) **mangeront**
 verbe
 indicatif, futur, 3ème personne, pluriel
- (2) **grand**
 masculin, singulier
 /grã/
 consonne latente = /d/
 ...

2.2. Information associée aux lexèmes

L'information associée aux *lexèmes* est de nature syntaxique (structure argumentale, traits sélectionnels, etc.) et sémantique (rôles thématiques des arguments, traits sémantiques, propriétés quantificationnelles, etc.). Elle comprend également une indication de la fréquence d'occurrence du lexème (toutes variantes morphologiques confondues). L'exemple (3) ci-dessous donne une illustration des informations associées au verbe *dire* :

- (3) **dire**
 [thème₁, but₂] : [__ NP₁ PP₂], [CP₁ PP₂]
 [thème₁] : [__ NP₁]
 ...

Relevons que dans (3), les différentes lectures syntaxiques du verbe *dire* sont représentées sous la forme de structures thématiques, par exemple [thème₁, but₂], où *thème* est le premier argument et *but* le second, avec l'indication de la réalisation syntaxique de ces arguments. Ainsi, dans notre exemple, le premier argument (thème) peut-il être réalisé soit comme un syntagme nominal, soit comme une proposition. Dans les deux cas, le deuxième argument est réalisé sous la forme d'un syntagme prépositionnel.

2.3. Information liée aux expressions

Nous considérons comme expressions idiomatiques des expressions à mots multiples dont la forme est figée ou semi-figée et/ou dont la sémantique n'est pas complètement compositionnelle comme *casser sa pipe* ou *rendre hommage*, que nous distinguons des mots composés comme *chauve-souris* ou *pomme de terre* (cf. Habert et Jacquemin (1993) ; M. Gross (1986) ; G. Gross (1990), pour un exposé des problèmes liés aux mots composés, Abeillé et Schabes (1989) ; Wasow *et al.* (1994), pour les expressions idiomatiques).

Une entrée de type *idiome* contient le lexème support de l'expression (le verbe

ou le substantif de base), un terme secondaire, la liste des constituants de l'expression, et une liste des contraintes associées à l'expression entière ou à ses constituants (passivisation, plurification, modification adjectivale, référents des possessifs, etc.). Le terme secondaire sert de clé de recherche secondaire, ce qui permet d'optimiser la recherche d'une expression. Cela est particulièrement utile dans le cas d'expressions basées sur un verbe support très fréquent, comme dans les expressions *casser sa pipe*, *donner sa langue au chat*, ou *faire le zouave*¹. Accessoirement, le terme secondaire permet d'établir facilement la liste de toutes les expressions contenant un terme donné. Par exemple, pour le mot *tombe*, *creuser sa tombe avec ses dents*, *être muet comme une tombe*, *avoir un pied dans la tombe*, etc. Un exemple d'entrée est donné en (4) :

- (4) *casser sa pipe*
 verbe support = casser
 terme secondaire = pipe
 [casser] [POSS pipe]
 [- passif], [POSS = sujet], ...

4. Rôle et place de la morphologie

La morphologie joue un double rôle, à la fois dynamique et statique, dans la base de données LMI. C'est tout d'abord une interface dynamique invoquée lors de l'insertion de nouveaux termes dans la base de données. En effet, pour garantir la cohérence dans les différents dictionnaires monolingues, les nouveaux termes ne peuvent être insérés que par l'intermédiaire de leur forme de référence (infinitif pour les verbes, singulier pour les substantifs, masculin singulier pour les adjectifs). Sur la base de spécifications initiales (catégorie, type de conjugaison ou de déclinaison) fournies au système par l'utilisateur, l'interface morphologique génère automatiquement ou de manière interactive un paradigme complet, qui doit être validé par l'utilisateur avant son insertion. Ce mode de vérification garantit la complétude et l'exactitude des entrées de LMI, à savoir, que toutes les formes fléchies d'un paradigme, et rien qu'elles, sont insérées dans la base de données. À côté du rôle dynamique que nous venons de discuter, la morphologie joue un rôle statique ou relationnel dans LMI, où elle prend la forme de liens entre entrées lexicales complètes. Ces liens expriment des relations flexionnelles, homographiques/homonymiques, ainsi que certaines relations d'ordre dérivationnel.

2.4.1. Relations flexionnelles

Comme l'illustre la figure 1, un paradigme flexionnel entier est relié à la liste des lexèmes qu'il réalise ; inversement, une liste de lexèmes est reliée à l'ensemble du paradigme de mots orthographiques. Ce double lien assure la correspondance entre mots et lexèmes aussi bien en analyse qu'en génération.

¹ Pour une discussion plus détaillée du problème de l'identification des expressions idiomatiques dans l'analyseur syntaxique du LATL, voir Campone et Wehrli (1996)

Mots	Lexèmes
go	
goes	go 1 (John has gone home)
going	go 2 (he is going to buy it)
gone	go 3 (he is going next)
went	...

FIGURE 1 : Relations morphologiques flexionnelles dans LMI.

2.4.2. Homographie/Homonymie

Lorsque des homographes sont associés à des paradigmes différents, comme c'est le cas pour l'exemple anglais illustré dans la figure 2, ces paradigmes sont enregistrés séparément dans le système.

Mots	Lexèmes
depression	[MEDICINE] (to suffer from depression)
depression	[METEOROLOGY] (a deep depression)
depressions	[ECONOMICS]

FIGURE 2 Homographie.

Cette façon de faire permet d'éviter certains pièges bien connus. Par exemple, considérons le cas de la traduction de la phrase (5), extraite d'un traité de médecine.

(5) Les dépressions sont plus fréquentes en automne qu'en été.

Comme il existe une entrée *dépression* portant le marqueur contextuel **MEDICINE** associée à un paradigme qui n'inclut aucune forme du pluriel, le système de génération pourra correctement produire la phrase (6a) plutôt que la forme incorrecte (6b).

- (6) a. Depression is more frequent in Falls than in Summer.
 b. *The depressions are more frequent in Falls than in Summer.

Si la phrase (5) provenait d'un rapport météorologique plutôt que d'un traité de médecine, c'est la traduction (6b) et non (6a) qui serait alors appropriée.

2.4.3. Relations dérivationnelles

Certaines relations d'ordre dérivationnel sont également représentées dans la base de données sous la forme de liens entre entrées lexicales. C'est le cas, notamment, de certains dérivés nominaux. Ainsi, des substantifs anglais comme *elaboration* ou *destruction* sont reliés aux verbes dont ils sont dérivés. Cette relation permet de récupérer facilement de l'information pertinente (typiquement, de l'information thématique pour les substantifs dérivés de verbes).

Lexèmes

destruction (N, ...)

destroy (V, [Agent, Thème], ...)

2.5. Traitement des mots composés

Dans les dictionnaires monolingues les mots composés courants (*rendez-vous*, *pillon de nuit*, *little by little*) apparaissent comme des entrées indépendantes des mots dont ils sont constitués. Rappelons qu'en général, la sémantique de ces mots composés n'est pas compositionnelle, ce qui signifie que leur sens ne peut être complètement dérivé du sens de leurs parties (p. ex. *zoot suit* « costume de zazou »).

En ce qui concerne l'allemand, langue dans laquelle le processus de composition de mots est particulièrement productif, plusieurs stratégies sont mises en œuvre en fonction de la nature du mot composé. Les verbes à particules séparables (comme *abfahren* « partir en véhicule » ou *weggehen* « partir à pied ») sont associés au verbe support (*fahren* et *gehen* respectivement), et chacune des combinaisons correspond à un lexème particulier (ou à une liste de lexèmes particuliers). Les substantifs composés lexicalisés, comme *Bahnhof* « gare », quant à eux, sont insérés directement dans le dictionnaire. Les mots composés qui ne seraient pas trouvés dans le dictionnaire font l'objet d'une analyse morphologique qui se termine avec succès lorsque toutes les parties du mot composé sont présentes dans le dictionnaire (*Anziehungskraft* « force de gravitation ») ou lorsqu'un préfixe est validé comme un préfixe séparable (*weg* dans *weggegangen* « parti »).

3. Les dictionnaires bilingues

Les dictionnaires bilingues de la base de données LMI font un usage crucial de la distinction entre formes orthographiques et lexèmes, puisque seuls ces derniers servent de base au transfert lexical (cf. Wehrli (1985), entre autres).

Il est aisé de montrer que le transfert lexical ne peut s'effectuer au niveau des formes orthographiques de surface. En effet, ce niveau de représentation lexicale encode des traits morphologiques tels que le cas, le genre, le nombre, le temps, etc., qui sont généralement propres à une langue et sont susceptibles de varier en fonction du contexte syntaxique dans lequel ils apparaissent. Ainsi, dans une langue aux cas morphologiquement réalisés comme l'allemand, la forme exacte d'un élément cible de type nominal ou adjectival ne peut être établie avant que certaines propriétés syntaxiques de ce constituant n'aient été déterminées, et en particulier sa fonction grammaticale. Cette dernière dépend de multiples facteurs liés aux propriétés lexicales du verbe gouverneur et aux éventuelles transformations grammaticales appliquées lors de la dérivation (transformation passive, etc.). Pour ces raisons, il est indispensable que le transfert lexical soit exprimé au niveau plus abstrait des lexèmes. Dans le processus de traduction, la sélection de la forme morphologiquement appropriée d'un lexème intervient à un stade relativement tardif de la dérivation de la phrase cible, lorsque tous les traits syntaxiques pertinents ont été identifiés.

Les dictionnaires bilingues spécifient l'ensemble des relations possibles entre les lexèmes de la langue source et ceux de la langue cible. Comme les relations entre unités lexicales de deux langues sont très fréquemment de type 1 à n, un dictionnaire bilingue doit également contenir des informations susceptibles d'aider la composante de transfert à sélectionner la traduction la plus appropriée. De telles informations doivent permettre de lever des ambiguïtés du type illustré dans les exemples suivants :

- | | | |
|-----|------------------------------|---|
| (7) | a. français : <i>temps</i> | anglais : <i>time, weather, ...</i> |
| | b. anglais : <i>time</i> | français : <i>temps, fois, ...</i> |
| | c. français : <i>mur</i> | allemand : <i>Mauer, Wand, ...</i> |
| | d. allemand : <i>stimmen</i> | français : <i>voter, accorder (musique)</i> |

Les lexèmes des langues source et cible sont reliés de façon complètement réversible, puisque chaque entrée dans le dictionnaire bilingue spécifie un lexème source et un lexème cible. Lorsqu'un lexème source peut correspondre à plus d'un lexème cible, le dictionnaire contient autant d'entrées que de correspondances, comme le montre la figure 4.

français	anglais
avoir (V, [__ NP])	have (V, [__ NP])
avoir besoin (V, [__ PP])	need (V, [__ NP])
avocat (N)	avocado (N)
avocat (N)	lawyer (N)
casser sa pipe (V)	kick the bucket (V)

FIGURE 4 : Exemples de correspondances bilingues.

Dans le but de résoudre des incompatibilités argumentales ou d'aider à la sélection de l'élément cible le plus approprié, il est nécessaire d'ajouter aux entrées de la figure 4 des informations supplémentaires, en particulier sur les correspondances d'arguments, le contexte d'utilisation ou le sous-langage pertinent, des descripteurs, ainsi que des informations statistiques. Quelques-uns de ces éléments sont décrits dans les sections ci-dessous.

3.1. Transfert d'arguments

Pour tous les éléments lexicaux susceptibles de sélectionner des arguments, comme les verbes, certains adjectifs et substantifs, il est nécessaire de spécifier dans le dictionnaire bilingue la façon dont les arguments du prédicat source correspondent aux arguments du prédicat cible. Cela permet de gérer les cas de non-correspondance comme celui illustré en (8) :

- | | |
|-----|---|
| (8) | a. Cet homme [vous] fournira [tous les renseignements dont vous avez besoin]. |
| | b. This man will provide [you] [with all the information you need]. |

Le verbe *fournir* dans la phrase (8a) sélectionne un objet direct et un objet indirect. Toutefois, dans la traduction de cette phrase, donnée en (8b), l'objet direct du

verbe *provide* correspond à l'objet indirect de *fournir*, et le complément prépositionnel [*with all the information you need*] à l'objet indirect du verbe français.

Les cas de non-correspondance ne sont pas limités à la réalisation syntaxique des arguments internes, comme pour l'exemple précédent. C'est un fait bien connu qu'un argument externe dans une construction source peut correspondre à un argument interne dans la construction cible. Par exemple, l'argument externe du verbe français *manquer* correspond à l'argument interne de *miss* en anglais ou de *vermissen* en allemand, comme les phrases (9) le montrent :

- (9) a. Héloïse manquait à Abélard.
b. Abélard missed Héloïse.
c. Abélard vermisste Héloïse.

Si les correspondances argumentales sont spécifiées dans le dictionnaire, la gestion de tels cas devient relativement simple. Dans la mesure où l'analyseur reconnaît la structure argumentale de la construction source, la composante de transfert utilise l'information de correspondance argumentale pour déterminer comment chaque argument doit être réalisé dans la langue cible.

3.2. Descripteurs

Les descripteurs sont des descriptions brèves (synonymes, définitions, paraphrases, etc.) qui permettent de distinguer les différentes lectures d'un terme source particulier (dans les cas d'homographie et de polysémie). Pour illustrer ce point, admettons que le dictionnaire bilingue français-anglais contienne les deux correspondances données dans la figure 4 pour le mot *avocat*. La première de ces correspondances (*avocat / avocado*) pourrait avoir pour descripteur **Fruit**, et la seconde (*avocat / lawyer*) le descripteur **Homme** ou **Homme de loi**. Ces descripteurs sont utilisés comme matériel de désambiguïsation par le système de traduction automatique interactif. Ainsi, pour le syntagme nominal *un avocat* dans la phrase (10a), l'utilisateur (non nécessairement anglophone) pourra sélectionner la bonne lecture sur la base d'un menu de dialogue du type (10b) :

- (10) a. Jean a besoin d'un avocat.
b. avocat 1. fruit
 2. homme de loi

3.3. Contextes et sous-langages

L'information contextuelle est une marque associée à des correspondances qui sont spécifiques à un sous-langage donné. Par exemple, le mot français *compositeur*, qui se traduit généralement en anglais par *composer*, devient *typesetter* dans le sous-langage de la typographie. Par conséquent, la correspondance entre *compositeur* et *typesetter* porte le trait contextuel **typography**. De même, la correspondance entre *accord* et *chord* porte le trait **music**, comme on peut le voir dans la figure 5.

français	anglais	contexte
compositeur	composer	standard
compositeur	typesetter	typography
accord	chord	music

4. LMI en chiffres

Pour conclure ce bref exposé sur la base de données LMI, donnons quelques chiffres qui montreront tout à la fois l'ampleur du travail déjà accompli et de celui qui reste à faire !

Les dictionnaires monolingues comptent approximativement 20 000 lexèmes pour l'allemand et pour le français, 45 000 pour l'anglais, ce qui correspond à plus de 160 000 formes orthographiques en français et en allemand, et environ 85 000 en anglais.

Pour les dictionnaires bilingues, nous disposons de près de 20 000 correspondances entre le français et l'anglais, plus de 10 000 entre l'anglais et l'allemand, alors que le dictionnaire français-allemand est en préparation.

Enfin, en ce qui concerne les expressions idiomatiques, elles n'ont pour l'instant été prises en considération que pour le français, et notre lexique en compte un peu plus de 2 000.

Acquisition semi-automatique du lexique

Evelyne VIEGAS et Sergei NIRENBURG

Computing Research Laboratory, New Mexico State University, Las Cruces, USA

• Abstract •

*In this paper, we present the process of lexical acquisition as we defined it to build **Spanlex**, a Spanish lexicon, for **Mikrokosmos**, a semantics-oriented machine translation (MT) system between Spanish and English. In this paper, we discuss the types of information which must be included in a computational lexicon for this and similar applications such as, for instance, text generation. We also present the acquisition methodology we developed which supports team acquisition!*

1. Introduction

Mikrokosmos est un système de traduction automatique entre l'espagnol et l'anglais, de textes journalistiques appartenant en priorité au sous-domaine terminologique d'opérations d'achat, de vente ou de fusion entre compagnies. Le système Mikrokosmos est basé sur la sémantique (Nirenburg *et al.*, 1994) et adopte une approche interlangue pour la traduction. La représentation interlangue est appelée une (TMR) *Text Meaning Representation* ou représentation de sens du texte. Nous ne pouvons développer, dans cet article, le processus complet de traduction ; nous nous contenterons de décrire très partiellement la représentation interlangue TMR, qui est essentiellement constituée des TMR provenant du lexique, où elles apparaissent à l'état non-saturé². Les lexèmes acquis proviennent d'un corpus de textes journalistiques qui présente un vocabulaire spécifique, mais aussi du langage courant, de par les produits dont il est question (voitures, pharmacies, ..). Ainsi, dans notre tâche d'acquisition

1. Toute notre gratitude à Victor Raskin, pour sa participation très active dans la mise en place du protocole d'acquisition. Nous remercions également tous nos lexicographes-acquéreurs pour leur acquisition effective, en particulier Oscar Cossio, Margarita Gorzales, Jeff Longwell, Maya, Javier Ochoa

2. Par « état non-saturé », nous entendons non-instancié. Par exemple, le mot *manger* donne des indications dans sa TMR sur ses contraintes de sélection : ANIMAL pour l'agent et MANGEABLE pour le thème ; ce n'est qu'au niveau de la TMR complète, lorsque toutes les TMRs venant d'autres lexèmes ont interagi, que l'on a une TMR saturée ; ainsi, dans *Jean aime manger chaud*, ANIMAL sera contraint à HUMAIN

lexicale, nous travaillons en collaboration étroite avec des lexicologues et des terminologues.

Notre approche théorique se situe dans une perspective de linguistique computationnelle qui privilégie la notion d'organisation lexicale, au sein d'une sémantique semi-compositionnelle. En d'autres termes, nous tirons avantage des techniques de l'intelligence artificielle et de la sémantique linguistique.

Notre but est d'acquérir un lexique d'environ 40 000 sens de mots. Cela rend nécessaire, si ce n'est inévitable, de semi-automatiser la tâche d'acquisition pour les tâches « rebutantes », comme, par exemple, vérifier l'orthographe d'un lexème de façon à ce que l'acquéreur puisse se concentrer sur des tâches plus productives et intéressantes (informations syntagmatiques, paradigmatiques, stylistiques ou pragmatiques).

Cela suggère en priorité d'avoir accès à des dictionnaires informatisés, à des corpus informatisés et surtout à des interfaces permettant une manipulation aisée de ces ressources. Nos interfaces ont été élaborées en accord avec l'utilisateur et continuent d'évoluer en fonction des besoins et des problèmes rencontrés lors de leurs utilisations.

Dans cette première étape de Mikrokosmos, nous avons essentiellement mis l'accent sur l'acquisition de l'information syntactico-sémantique, effectuée par le biais de dépendances sémantico-syntaxiques.

Nous présentons à l'acquéreur une série de moules sémantico-syntaxiques pré-définis, le guidant dans la phase d'acquisition. Lexicographes et terminologues utilisent le même outil d'acquisition, puisqu'il leur est possible de spécifier différents types d'informations (le domaine d'application, le lieu d'emploi, les collocations, etc.).

Dans ce qui suit, nous présentons tout d'abord le type d'information que l'on trouve dans les lexiques computationnels et la façon dont cette information est structurée et organisée. Puis, nous motivons et présentons le type d'information que nous codons dans nos propres lexiques, en donnant l'exemple de **Spanlex**, le lexique espagnol acquis dans le cadre du projet Mikrokosmos. Nous passons ensuite à la tâche même d'acquisition et présentons les différentes étapes intervenant dans ce processus, en montrant qu'il est possible d'acquérir un lexique de haute qualité et à grande échelle.

2. Quels types de lexiques pour quels types d'application : la question est-elle vraiment pertinente ?

Les principaux lexiques computationnels que l'on trouve sur le réseau ou qui sont en cours de développement à l'heure actuelle, renferment essentiellement deux types d'information : essentiellement syntaxique, comme *Comlex*, (Macleod et Grishman, 1994) et/ou sémantique, par exemple *Acquilex*, (Sanfillippo, 1992). Un autre type de différences entre ces deux types de lexiques, se situe au niveau de la représentation, qui adopte soit une hiérarchie lexicale, où les feuilles de la hiérarchie sont des lexèmes ; ou qui adopte une hiérarchie conceptuelle, où les lexèmes sont reliés entre eux via les concepts. C'est cette dernière approche que nous adoptons dans Mikrokosmos.

En ce qui concerne l'organisation globale du lexique et quels lexèmes vont donner lieu à une entrée dans le lexique computationnel, il est à noter que les lexiques qui mettent l'accent sur la syntaxe adoptent une approche par énumération de sens basée sur des sous-catégorisations différentes : *vouloir* et *vouloir que...*, reçoivent des entrées différentes, comme nous l'expliquons dans le paragraphe suivant.

2.1. Motivation de la description de nos lexiques

Le type d'information contenu dans un dictionnaire dépend très souvent du **type d'application** pour lequel il est utilisé. Par exemple, pour faire de la traduction multilingue, des dictionnaires multilingues, où l'on juxtapose les lexèmes des différentes langues, peuvent être suffisants : *manger/comer/eat*, respectivement en français, espagnol et anglais. Pour faire de la génération, de l'information sur l'ordre des mots, par exemple la place de l'adjectif dans un groupe nominal, *une maison bleue* et non *une bleue maison*, est nécessaire ; de même, il est nécessaire de coder les collocations dans un dictionnaire, *a heavy smoker* versus *un grand fumeur*, respectivement en anglais et français. Il est clair que l'information collocationnelle n'est pas indispensable en phase d'analyse.

Cependant, l'acquisition de dictionnaires à grande échelle est un travail coûteux, c'est pourquoi il est préférable d'acquérir un lexique qui soit réutilisable pour d'autres domaines, d'autres applications.

Cela nous amène à nous concentrer maintenant sur le type d'**organisation et de structuration du lexique**. Il est bien établi, à l'heure actuelle, en sémantique lexicale computationnelle, qu'une simple énumération des mots par ordre alphabétique est computationnellement chère (Boguraev et Pustejovsky, 1990). Par ailleurs, il est également reconnu qu'une structuration du sens des mots basée sur des sous-catégorisations différentes est également chère, computationnellement parlant, et surtout inadéquate, en ce qu'elle empêche de capter le noyau sémantique de l'item lexical, par exemple *oublier* qui sous-catégorise pour un groupe nominal dans *J'ai oublié les clés de ma voiture* ou un xcomp dans *j'ai oublié de prendre mes clés* ou encore un comp dans *j'ai oublié que tu avais mes clés*, (Viegas et Nirenburg, 1995). Les lexiques que nous construisons intègrent les résultats les plus avancés, qu'ils proviennent de la sémantique lexicale (distinction de sens de mots), de la linguistique computationnelle (dépendances sémantico-syntaxiques), ou de l'intelligence artificielle (en termes de techniques, comme les hiérarchies à héritage simple ou multiple). La construction de ce type de lexiques implique de faire appel à des experts en lexicographie, terminologie, linguistique computationnelle, et représentation des connaissances.

Dans ce qui suit, nous donnons une description documentée de l'information contenue dans nos lexiques. Le lexique est composé de super-entrées (suivant la convention adoptée par Meyer *et al.*, 1990). Chaque super-entrée consiste en une liste d'entrées représentant des sens de mots différents, et ce indépendamment de la catégorie syntaxique du mot. Chaque sens de mot est identifié par un unique identificateur, ou un lexème (suivant la terminologie de (Mel'čuk *et al.*, 1984)). À l'intérieur d'une entrée on peut trouver des noms, verbes (en anglais, *walk-N* versus *walk-V*), ou des homonymes. Les critères utilisés pour décider ou non de la création d'une

nouvelle entrée ou d'une sous-entrée sont les suivants (Onyshkevych et Nirenburg, 1994) :

1 Candidats potentiels à une super-entrée :

- les noms composés représentés par un seul mot ou séparé par un tiret (rouge-gorge), sont des candidats potentiels ;
- les noms propres, composés ou non, sont stockés dans une base de connaissances, autre que celle du lexique de base ; ils sont néanmoins codés en utilisant le même format ;
- les idiomes sont indexés sur la tête, par exemple, *battre* dans *battre le fer (tant qu'il est chaud)*.

2 Critères de sélection pour un nouveau lexème :

Nous renvoyons à (Meyer *et al.*, 1990) pour la justification des critères linguistiques et lexicographiques qui permettent de décider la création ou non d'une nouvelle sous-entrée.

Concrètement, en phase d'acquisition, la façon dont est structurée notre ontologie, joue un rôle primordial dans la façon d'acquérir un lexème. Si nous reprenons l'exemple de *livre*, qui est ambigu entre l'ASPECT PHYSIQUE et l'ASPECT DOCUMENT, la décision de créer une entrée ou deux est « suggérée » au lexicographe, dans la mesure où l'ontologie prévoit les deux types pour *livre*, comme le montre la figure ci-dessous, (Kavi, en préparation).

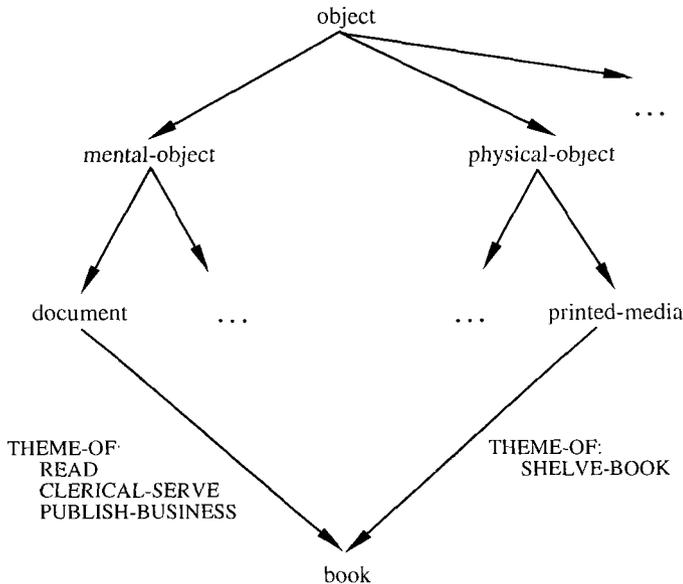


FIGURE 1 : Ontologie pour livre.

2.2. Les différentes zones d'un lexème

L'information contenue dans un lexème est répartie selon des **zones** correspondant à divers niveaux d'information lexicale, que nous décrivons ci-dessous (Meyer *et al.*, 1990).

- 1 **CAT**égorie syntaxique : *Nom, Verbe, Pronom*
- 2 **MORPH**ologie : pour les formes irrégulières et les changements de stems *mouse* versus *mice* en anglais.
- 3 **COMMENTS** : cette zone est subdivisée en plusieurs zones donnant de l'information administrative (date d'entrée du système pour l'acquisition du lexème, date de modification, nom du lexicographe), une définition du mot, des exemples dans la langue source, le domaine d'application du mot (langue, région).
- 4 **ORTH**ographe : pour les abréviations ou variantes *United States of America* versus *USA*.
- 5 **PHON**ologie
- 6 **SYN**tactic-**STRUC**ture : donne des indications sur les dépendances syntaxiques au niveau de la proposition ou de la phrase ; cette zone renferme essentiellement de l'information sous-catégorielle.
- 7 **SEM**antic-**STRUC**ture : la partie sémantique lexicale du mot, donnant sa TMR non saturée, ou représentation du sens.
- 8 **LEX**ical-**REL**ations, donne de l'information collocationnelle.
- 9 **LEX**ical-**RULES** : répertorie l'ensemble des règles qui s'appliquent à ce lexème.
- 10 **STYL**istique : donne de l'information sur les facteurs stylistiques, tels que le degré de familiarité, de formalité. Cette zone contient aussi des sous-zones contenant des fonctions « déclencheuses » pour l'analyse (pour traiter la co-référence) ou pour la génération (donnant la préférence au niveau de l'ordre des mots, par exemple.)

2.3. Une base de connaissances multi-propos

De la description précédente, par rapport au type d'information codée dans le lexique, il découle que nos lexiques sont multi-propos, en étant conformes aux trois points ci-dessous :

- a **multi-lingues** : ils acceptent plusieurs langues naturelles aussi différentes que l'espagnol et le japonais,
- b **multi-media** : ils renferment de l'information linguistique, pour le traitement du langage naturel, de l'information phonologique, essentiellement pour un traitement de reconnaissance de la parole, et enfin de l'information pour la vision, pour une reconnaissance visuelle, par le biais de l'ontologie.
- c **multi-usages** : ils peuvent être utilisés en analyse, et génération monolingue ou multilingue ; en traduction (semi)-automatique ; pour la reconnaissance/production de la parole.

3. Le processus d'acquisition : les différentes tâches

Nous consacrons les paragraphes suivants au processus même d'acquisition de la connaissance syntactico-sémantique, tel que nous l'avons développé pour **Spanlex**. Le schéma ci-dessous représente tous les modules de travail et de ressources nécessaires à l'acquisition du lexique décrit ci-dessus. Dans ce schéma figurent également les outils de test/évaluation des entrées lexicales, ainsi que l'analyseur sémantique qui utilise dynamiquement toutes les ressources statiques (lexique, ontologie).

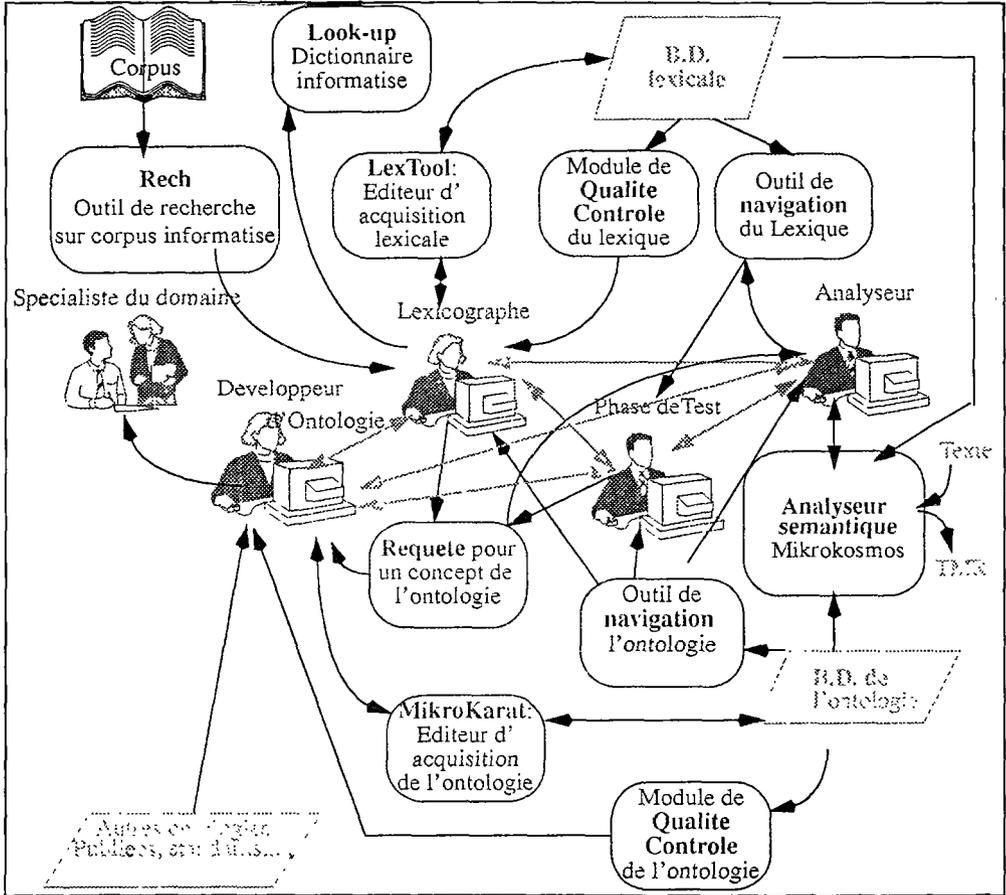


FIGURE 2. Processus d'acquisition : modules de travail et ressources.

Il est important de noter le nombre d'interactions entre les « développeurs » de lexiques et les « développeurs » d'ontologies. En effet, l'ontologie n'est pas figée une fois pour toutes, elle évolue en fonction des lexèmes rencontrés dans différentes langues. Pour cela, nous avons développé des outils de communication où des requêtes aux « développeurs » d'ontologie sont formulées périodiquement.

4. Quelques chiffres explicatifs

Nous nous concentrons maintenant sur le processus d'acquisition tel que nous l'avons développé pour Spanlex, le lexique espagnol que nous sommes en train de construire. Notre tâche consiste à acquérir de 30 000 à 40 000 sens de mots espagnols. Le type d'information que nous codons dans le lexique, à savoir essentiellement lexicosémantique, ne peut se faire si ce n'est de façon semi-automatique. Pour cela, nous avons conçu et implémenté des outils d'aide à l'acquisition proprement dite (voir figures en Annexe 1), de façon à faciliter la tâche aux lexicographes, ainsi que des outils d'évaluation de l'acquisition, de façon à pallier les inconvénients dus à une intervention humaine, pour une application computationnelle.

Nous décrivons ci-dessous le développement du processus sur une période de douze mois, de novembre 1994, date de l'initialisation de l'acquisition lexicale jusqu'à décembre 1995. Nous montrons comment il est possible effectivement de réaliser une acquisition lexicale à grande échelle, en incorporant essentiellement de l'information sémantique.

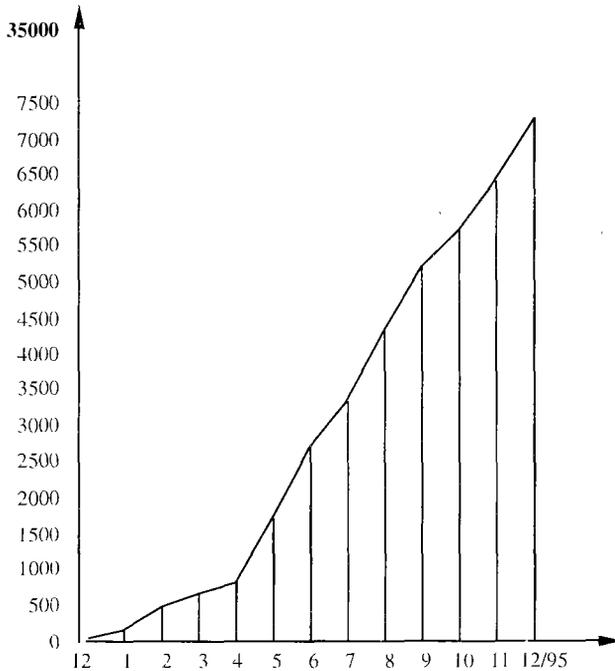


FIGURE 3 : Acquisition semi-automatique du lexique

La figure 3 montre l'évolution du nombre d'entrées sur les douze mois. Les chiffres représentés ici, indiquent le nombre de sens de mots ou de lexèmes acquis semi-automatiquement. Le premier travail a consisté en une analyse linguistique de façon à déterminer les formats pré-définis renfermant de l'information sous-catégorielle et les rôles thématiques associés, comme le format Trs-Ag-Th qui spécifie que le verbe est transitif et sous-catégorise pour un sujet et un objet qui ont les rôles respectivement

d'agent et de thème, par exemple le verbe *manger*. Ces formats, résultats d'une étude linguistique, ont été mis en place de façon à guider le lexicographe dans sa tâche d'acquisition. Nous avons également conçu une interface d'acquisition, qui a été élaborée en suivant le processus d'acquisition tel que les lexicographes le perçoivent et non pas en suivant la structure de la base de données. En plus de l'interface d'acquisition elle-même, nous avons mis à la disposition de nos lexicographes des dictionnaires monolingues espagnols et bilingues, espagnol-anglais. Par ailleurs, un outil de recherche de mot en contexte, dans un corpus de textes espagnols en ligne a été également créé (figure en Annexe 2). Ces outils sont décrits plus amplement dans (Viegas, 1995). Ce type d'acquisition requiert un important effort d'apprentissage de la part des lexicographes, et est aussi coûteux. Cependant, si l'on veut constituer un lexique syntactico-sémantique de base, cet effort est inévitable.

Nous avons développé en parallèle un programme produisant entièrement automatiquement les entrées, attestées par les dictionnaires et/ou corpus, des nouvelles formes dérivationnelles de lexèmes acquis semi-automatiquement, et ce à l'aide de règles morpho-sémantiques.

Voici un exemple partiel, des formes générées automatiquement, avec une règle sémantique associée, pour le verbe espagnol *comprar* (acheter) :

comprar, v, LR1event
comprador, n, LR2social_role_relation1a
compra, n, LR2event10
compra, n, LR2theme_of_event10
compradero, adj, LR3feasibility_attribute2a
comprable, adj, LR3feasibility_attribute1
comprado, adj, LR3event_telic
compradizo, adj, LR3feasibility_attribute5a
comprador, adj, LR3social_role_relation1a
malcomprar, v, LRneg_affect1, LR1event
malcomprado, adj, LR3event_telic
recomprar, v, LRrepetition1, LR1event
recompra, n, LR2event10
recompra, n, LR2theme_of_event10
recomprado, adj, LR3event_telic

Par exemple, *comprable*, adj, LR3feasibility_attribute1, est dérivé morphologiquement de *comprar*, et ajoute à la sémantique de *comprar* la caractéristique d'être possible ou non.

Les nouvelles entrées générées, sont ensuite testées par les lexicographes, à l'aide des mêmes outils qui testent les entrées acquises semi-automatiquement ; nous allons pouvoir ainsi multiplier la taille de notre lexique³, par 5 ou 6 (figure 4), et ce automatiquement (Viegas et Gonzales, 1995).

3. Nous avons commencé la génération morpho-sémantique à partir des 1 500 verbes déjà acquis, et le nombre moyen de formes lexicales générées par verbe est de 35.

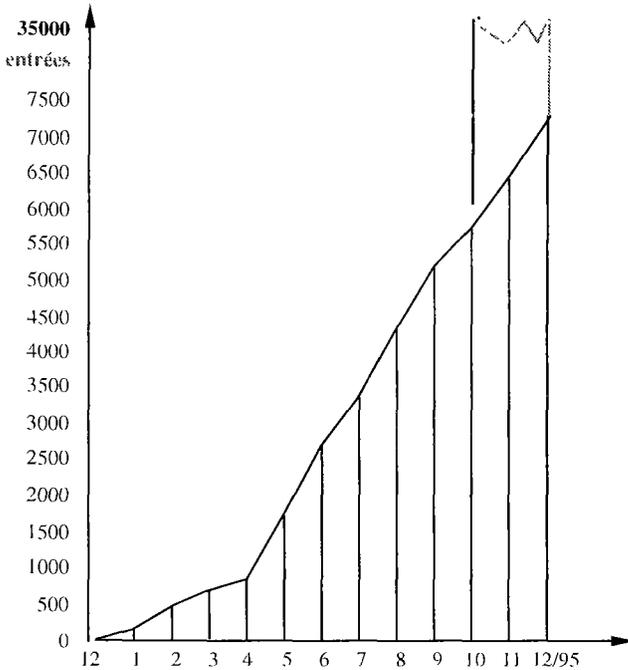


FIGURE 4 Développement du lexique à partir de règles morphologiques dérivationnelles.

Dans ce processus de génération automatique, le but est de pouvoir générer automatiquement la sémantique des dérivations morphologiques des lexèmes déjà acquis, afin de les présenter à l'acquéreur pour vérification, tout en allégeant la tâche de l'acquisition manuelle, et tout en enrichissant notre lexique⁴.

5. Conclusion

Dans cet article, nous avons présenté les ressources nécessaires pour réaliser l'acquisition semi-automatique de lexiques riches en information sémantique. Nous avons également mis l'accent sur, l'importance d'une intervention humaine d'une part, et d'autre part sur l'importance d'avoir des outils pour guider l'acquisition, et des outils pour évaluer les résultats de l'acquisition, de façon à élaborer des lexiques de haute qualité et utilisables pour un traitement (semi-)automatique du langage naturel. Il est aussi primordial, d'aller vers la construction de bases de connaissances lexicales qui puissent être utilisées par différentes applications ou dans différents domaines.

4. Viegas et Nirenburg (+996) expliquent en détail le processus de génération morpho-sémantique.

Annexe 1 : Outils d'acquisition

Word to find:

Files to search: /home/mikro/evelyne/lexicon/esp */

Verb forms

Noun forms

Partial word

LOAD TEMPORARY WORK

LOAD LIST OF WORDS

***** LEXICON BUILDER INTERFACE FOR PART OF SPEECH *****

0 3 0 0 0 0 0 0 0 0 0 0

VERB
 NOUN
 ADJECTIVE
 ADVERB
 PREPOSITION
 PRONOUN
 DETERMINER
 CONJUNCTION
 POSSESSIVE
 DEMONSTRATIVE
 REFLEXIVE
 NEGATION

UNIV-OBJ-MAP
 UNIV-EVENT-MAP
 CONSTRAIN-MAP

Constrain-Event-Obt
 Constrain-Event-Mult
 Constrain-Property-Ob
 Reassign-Mapping
 Constrain-Map-Autind

***** LEXICON BUILDER INTERFACE FOR PART OF SPEECH *****

(CAT n)
 (MORPH
 (ANNO

(SYN-STRUC
 (root \$vcat0)
 (cat n)
 toblique (root \$var)

(SEM-STRUC
 (LEX-MAP

(\$var) (n)
 (n) (n)
 (\$Svar) (ms)

(DEF "the imparting of thoughts, opinions")
 (EX "Dieron a conocer sus planes mediante
 comu.n.aci.on")
 (COMMENTS Event-Comm
 (TIME-STAMP 11 21 33 Jan 57 11 07 21")
 (LAST-STOP)
 (CROSS-REF)
 (SYN-STRUC
 (1 (root \$vcat0)
 (cat n)))
 (SEM-STRUC
 (LEX-MAP
 "statement"
 (1 (declin
 (aspect ((telic, yes)))))
 (LEX-RULES
 (SYN-GAT
 (SYN-ANNO
 (communication-N2
 (CAT n)
 (MORPH
 (ANNO
 (DEF "document or message imparting information")
 (EX "Dieron la comunicaci.on del presidente")
 (COMMENTS Object-Comm
 (TIME-STAMP 11 23 11")
 (LAST-STOP)
 (CROSS-REF)
 (SYN-STRUC
 (1 (root \$vcat0)
 (cat n)))
 (SEM-STRUC
 (LEX-MAP
 "message"
 (1 (document)))
 (LEX-RULES

LOAD TEMPORARY WORK

LOAD LIST OF WORDS

***** LEXICON BUILDER INTERFACE FOR PART OF SPEECH *****

0 3 0 0 0 0 0 0 0 0 0 0

VERB
 NOUN
 ADJECTIVE
 ADVERB
 PREPOSITION
 PRONOUN
 DETERMINER
 CONJUNCTION
 POSSESSIVE
 DEMONSTRATIVE
 REFLEXIVE
 NEGATION

UNIV-OBJ-MAP
 UNIV-EVENT-MAP
 CONSTRAIN-MAP

Constrain-Event-Obt
 Constrain-Event-Mult
 Constrain-Property-Ob
 Reassign-Mapping
 Constrain-Map-Autind

***** LEXICON BUILDER INTERFACE FOR PART OF SPEECH *****

JS-A
 VALUE LANGUAGE-RELATED-OBJECT

SUBCLASSES: CONTRACT BOOK BALLOT ARTICLE ACADEMIC PAPER LEDGER TEXT MANUAL LIST REPORT PLAN

PART OF : OBJECT

DEFINITION
 VALUE anything printed or hand-written that is relied upon or used as proof of something

PRODUCED-BY
 SEM HUMAN

Notes Show Show Page Help Unit

346

Annexe 2 : Recherche d'occurrences dans le corpus

Word to find:

Files to search:

Verb forms

Noun forms

Partial word

La **comunicación** oficial no precisó el monto de la operación tele**comunicación**es y Hachette (edición y comunicación) se la **comunicación** presentada por el comisario europeo encargado de las **comunicación** de orientación política, ya que la **comunicación** de la Comisión Europea será presentada al **comunicación** y el niquel.

publicación servicio permanente a los medios de **comunicación** **comunicación** y que se **comunicación** sea más operativo alquiler de líneas de **comunicación** vía satélite en las **comunicación** la región acceder a las grandes autopistas de la **comunicación** como entre el Atlántico y el Pacífico, con apenas **comunicación** en un intento por aceptar el mercado de la **comunicación** en EEUU, en un intento por aceptar el mercado de la **comunicación** en EEUU, en un intento por aceptar el mercado de la **comunicación** en medios de **comunicación** de habla hispana **comunicación** social competitividad en el mercado hispano de la **comunicación** lujo en la principal vía de **comunicación** que une las dos ciudades informó hoy la Secretaría de **comunicación**.

EEUU-COMUNICACIONES (previsto: **comunicación**es- AT&T (American Tele- **comunicación**es celulares y radiadas System. Aquella decisión no trató al **comunicación**es Comisión Federal de **comunicación**es de

Find

Show Forms

Viewing /home/mikro/evelyne/lexicon/esp merg-acq/eval93-1.esp
laboratorio farmacéutico Doctor Andreu, se informó hoy aquí.

La **comunicación** oficial no precisó el monto de la operación realizada entre Productos Roche SA y Unión Explosivos Rio Tinto SA, hasta ahora mayoritaria en el accionariado

Fuentes financieras consultadas citaron la operación en unos 10 000 millones de pesetas. Según el acuerdo firmado hoy en Madrid los productos del Doctor Andreu continuarán siendo

Next | Prev

Viewing /home/mikro/evelyne/lexicon/esp merg-acq/mergers_n_acquisitions 122text
con México, aunque teniendo en cuenta "la sensibilidad de ciertos productos", supo EFE de fuentes comunitarias.

La **comunicación** presentada por el comisario europeo encargado de las Relaciones con América Latina, Manuel Marín, recibió hoy, lunes, luz verde en la reunión preparatoria de jefes de gabinete de los comisarios, lo que permitirá su aprobación como punto A (sin debate) el miércoles

La propuesta de Marín, que ha tenido muy en cuenta la pertenencia de México al Tratado de Libre Comercio norteamericano

Next | Prev

comunicación

comunicaciones

Dismiss

347

