IDAREX : description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis*

Frédérique SEGOND et Élisabeth BREIDT

Rank Xerox Research Centre, Meylan, France, et Université Tübingen, Allemagne

1. Introduction

La plupart des textes regorgent d'expressions à mots multiples. Ces expressions ne peuvent être correctement comprises, encore moins traduites, si elles ne sont pas reconnues en tant qu'unités lexicales complexes. Ces expressions, appelons-les **lexèmes à mots multiples** (LMM), englobent aussi bien les expressions idiomatiques (*se casser la tête sur quelque chose*), les proverbes (*Bien mal acquis ne profite jamais*), les constructions à verbe support (*prendre la parole*) que les collocations lexicales et grammaticales (*au vu de*). Nous présentons ici une méthode permettant de repérer de telles expressions dans un texte. Ce travail est fait pour le français et pour l'allemand. D'abord nous illustrons à l'aide d'exemples comment les LMMs, tout en obéissant à la syntaxe générale, ont un certain nombre de restrictions qui leur sont attachées en fonction des variations lexicales et/ou structurelles qu'ils autorisent : comment les LMMs sont régis par une syntaxe qui leur est propre. Nous nous proposons ensuite d'encoder le comportement intrinsèque des LMMs dans des **règles de grammaire locales**. L'implantation de ces grammaires locales est réalisée dans le cadre général des techniques à états finis (Karttunen et Yampol, 1993 et Karttunen et Beesley, 1992) à l'aide du formalisme à deux niveaux IDAREX (Segond et Tapanainen, 1995 et Tapanainen, 1994). Cet environnement, qui s'avère rapide et efficace, permet d'encoder les LMMs comme des expressions régulières. Ces règles de grammaires locales ont été utilisées dans le cadre du projet COMPASS (Adaptation de dictionnaires bilingues pour l'aide à la compréhension en ligne)[1].

* Nous tenons à remercier Jean-Pierre Chanod et Alain Lecomte pour leurs commentaires.
1. Projet de recherche LRE 62-080, financé par la CEE

2. Variabilité des LMMS

Certaines LMMs apparaissent toujours dans une forme donnée et sont donc facilement identifiables par leurs seuls éléments lexicaux. C'est le cas de *de fil en aiguille* ou de *par acquis de conscience*. Cependant la plupart des LMMs autorisent différents types de variations et de modifications[2]. Pour reconnaître de tels LMMs dans les textes il faut être en mesure d'identifier précisément celles de leurs occurrences qui dévient de la forme standard ou forme de base. Ainsi il faut être attentif aux différentes variations morphologiques, à l'ordre des mots ou encore à l'insertion de composants. Par exemple, si l'on considère l'expression « *casser sa pipe* », on s'aperçoit que le nom *pipe* ne peut être mis au pluriel, que le verbe *casser* ne peut être remplacé par un de ses synonymes (par exemple *briser*), et que la phrase ne peut être mise au passif sans que le sens de l'expression idiomatique soit perdu. Pourtant le verbe lui-même peut être conjugué. De même l'expression allemande *die Beine in die Hand nehmen* (prendre ses jambes à son cou) n'autorise aucune variation lexicale, le nom *Hand* ne peut être mis au pluriel, enfin la phrase ne peut être ni topicalisée, ni passivisée. Les exemples qui suivent montrent que l'expression idiomatique *sich über etwas den Kopf zerbrechen* (se casser la tête sur quelque chose) admet plusieurs variations. Les points de suspension indiquent les positions où l'insertion d'adjoints est possible.

Über diese Sache [...] *zerbrach* Jan *sich* schon lange [...] *den Kopf.*
Deswegen zerbrach sich Jan schon lange [...] *den Kopf.*
Sich darüber [...] *den Kopf* zu *zerbrechen*, lohnt sich nicht.
Jan *zerbricht sich* selten [...] *den Kopf über* solche Dinge.
Jan *zerbricht sich* nur *über* wenige Dinge im Leben [...] *den Kopf.*

De même on peut insérer un adverbe dans l'expression française *peser dans la balance* et obtenir ainsi les phrases suivantes : *peser lourd dans la balance* ou *peser beaucoup dans la balance*. En revanche l'insertion d'un GN dans cette phrase entraîne la perte du sens idiomatique : *peser les fruits dans la balance*. Le même type de phénomène se passe lorsqu'une expression idiomatique est incluse dans une phrase plus grande. L'expression *tomber du ciel* garde son sens idiomatique dans le cas d'une insertion d'adverbe (*ça tombe vraiment du ciel*) mais l'interprétation idiomatique échoue si l'on insert un GN comme dans la construction causative : *C'est la condensation qui fait tomber la pluie du ciel.*

Les types les plus communs de variations et de modifications pour lesquels nous donnons des exemples en section 4, sont les suivants[3] :

- **variations lexicales**
  variant lexical ; variation dans la réalisation des arguments ; échange des constituants ;

- **variations morpho-syntaxiques**
  variation en nombre ; variation casuelle (LMMs nominaux) (toujours possible en

2 Par *variations* nous entendons un échange ou une restructuration syntaxique des composants , par *modifications* nous entendons l'ajout de mots, de modifieurs, aux LMMs.
3 Voir également Fleischer (1982), Brundage *et al* (1992) et Engelke (1994)

allemand) ; forme comparative ou superlative pour un adjectif ; variation du déterminant ; variation de la personne du verbe ; variation du temps du verbe ; composition d'un élément nominal ;

- **modifications**
  modification adverbiale ; modification adjectivale ; négation ;

- **variations structurelles**
  passivisation ; topicalisation ; scrambling ; ordre des mots pour VI/V2/V finaux.

Sauf en ce qui concerne la variation de l'ordre des mots, les variations structurelles sont extrêmement restreintes. De même, la composition nominale et l'échange des constituants sont pratiquement toujours impossible. Nous supposons donc par défaut qu'aucune des variations citées précédemment n'est possible et que nous pouvons explicitement lister les exceptions. Les autres variations, telles que le nombre et le cas des noms et des adjectifs, la personne et le temps des verbes sont en revanche très courantes et seront aussi exprimées simplement dans le formalisme.

En plus de la variabilité qui doit être prise en compte dans la description des LMMs, il apparaît que certains LMMs enfreignent les règles d'accord et de gouvernement (valence du verbe déviante, nom quantifiable sans déterminant, adjectif ou préposition sans nom qui lui soit attaché). Une telle caractéristique peut être prise en compte, mais n'est pas nécessaire pour reconnaître les LMMs dans un texte.

Les exemples précédents font apparaître clairement l'insuffisance des méthodes de reconnaissance de patrons pour le repérage des LMMs. En effet, les LMMs autorisent des variations qui, pour la plupart, sont mal définies au niveau lexicographique. Une entrée classique de dictionnaire fournit généralement une seule forme pour l'expression – pas nécessairement la forme de base ou canonique –, sans aucune autre précision quant aux variations autorisées, mis à part, quelquefois, pour les variantes lexicales. Or cette information peut être encodée dans des règles de grammaire locales, règles qui ont un pouvoir d'expression plus grand que les descriptions traditionnelles.

Si on les compare à des règles de grammaire générales, les règles de grammaire locales décrivent implicitement les restrictions des LMMs. Ces restrictions énoncent les variations autorisées pour les LMMs par comparaison au cas, par défaut, où ils sont complètement figés. Dans le cas par défaut, toutes les restrictions s'appliquent, *i.e.* aucune variation n'est permise et le LMM est représenté par sa forme de surface dans laquelle tous les composants sont figés et ordonnés. Les violations des règles de grammaire standard, p. ex. constituants manquants ou accords enfreints, n'ont pas besoin d'être explicitées. Cependant elles peuvent être décrites lorsqu'elles importent pour la distinction entre sens idiomatique et sens littéral d'un LMM.

D'abord nous donnons brièvement le formalisme utilisé. Ensuite, nous décrivons, par le menu, comment sont exprimés dans les grammaires locales, les différents types de variations des LMMs.

3. IDAREX : Formalisme pour décrire la variabilité des LMMs

Les règles de grammaire locales sont écrites à l'aide du formalisme à deux niveaux

IDAREX[4] partie intégrante du compilateur à états finis développé au centre de recherches de Rank Xerox[5].

Les LMMs sont codés comme des expressions régulières en accord avec les notations décrites ci-dessous.

3.1. Les mots

Les mots sont représentés à deux niveaux : un niveau lexical et un niveau de surface. Les deux points servent de séparateur entre les deux niveaux. Il y a quatre descriptions de base possibles pour un mot.

1. :forme-de-surface
2. :forme-de-surface variable morphologique:
3. forme-de-base variable morphologique:
4. variable-classe-de-mot

Les deux premiers cas permettent la description des formes figées. Par exemple, la forme figée *pédales* dans l'expression *perdre les pédales* peut être encodée de deux manières :

1. *:pédales*
2. *:pédales Noun:*

Le troisième cas permet la description des formes variables. Dans l'exemple précédent le verbe *perdre* peut apparaître à n'importe quels temps, nombre et personne. Il est donc codé *perdre Verb:* ou la variable Verb stipule que toutes les réalisations verbales du mot *perdre* sont autorisées.

Le dernier cas permet la description de variables générales représentant des classes de mots. Ainsi nous pouvons par exemple définir une variable ADV regroupant les adverbes et les expressions adverbiales. Nous sommes maintenant en mesure d'écrire la règle de grammaire locale complète pour l'expression idiomatique précédente :

perdre Verb: ADV :les :pédales;*

3.2. Les opérateurs

Un ensemble d'opérateurs permet de combiner entre elles les descriptions des mots. Parmi eux :

- *rien* — les mots se succèdent les uns aux autres ;
- *parenthèses* () — marquent une partie optionnelle de l'expression idiomatique ;

4. **ID**ioms **A**s **R**egular **Ex**pressions.
5 Pour une description plus détaillée du formalisme nous renvoyons le lecteur à Karttunen et Yampol (1993) et Segond et Tapanainen (1995).

- *étoile de Kleene* * — marque que la chaîne qui la précède peut apparaître n'importe quel nombre de fois, y compris aucune ;
- *plus* + — marque que la chaîne qui la précède peut apparaître une ou plusieurs fois ;
- *crochets* [] — groupent une expression ;
- *barre* | — sépare les différentes possibilités ;
- *point virgule* ; — marque la fin d'une expression.

3.3. Les variables fonctionnelles

Enfin, le formalisme permet l'utilisation de variables fonctionnelles ou macros. Les lexicographes ont ainsi la possibilité de décrire, de façon compacte, des phénomènes complexes réguliers. Ces descriptions sont ensuite réécrites par le système. Nous avons utilisé des macros pour décrire, en français, les LMMs dans lesquels apparaissaient des pronoms réflexifs. Par exemple, pour décrire toutes les variations possibles, tout particulièrement les variations temporelles, de l'expression *s'aplatir comme une carpette devant quelqu'un.* Au lieu d'écrire :

REFL [être Verb: ADV* aplatir VPP: | aplatir Verb:] ADV* :comme :une :carpette (:devant NP)

nous écrivons simplement :

REFLEX(aplatir) :comme :une :carpette (:devant NP)

où `REFLEX(verbe)` est utilisé chaque fois qu'un LMM contient une construction réflexive et se réécrit en :

REFL [être Verb: ADV* verbe VPP: | verbe Verb:] ADV*

4. Règles de grammaire locales pour le français et l'allemand

Les règles de grammaire locales que nous proposons ici couvrent au maximum le niveau de la phrase. Elles sont énoncées de la façon la plus générale possible et autorisent la surgénération. Bien que des règles plus restrictives et plus spécifiques puissent être écrites elles ne sont pas nécessaires dans la mesure où l'on fait comme hypothèse de départ qu'il n'y a pas de phrase mal formée en entrée. En effet, peu importe que les règles autorisent plus de variations que celles qui apparaîtront effectivement dans les textes dans la mesure où l'on peut distinguer les emplois idiomatiques des emplois littéraux. Par exemple, étant donné que nous ne prenons pas en compte la représentation sémantique des LMMs. La règle de grammaire locale associée à l'expression française *peser dans la balance* acceptera aussi bien les phrases sémantiquement correctes comme *peser lourd dans la balance* ou *peser énormément dans la balance* que celle difficilement acceptables sémantiquement comme *peser *ardemment dans la balance.*

Certains LMMs ont une variabilité tellement productive qu'il est illusoire de vouloir en rendre compte de façon systématique. De telles variations sont, par leur na-

ture même, imprévisibles. Un exemple de cette productivité est donné par la formation *ad hoc* de mots composés ou par la combinaison de métaphores et d'expressions idiomatiques en allemand comme dans :

das bißchen Kopf, das sie noch haben, zerbrechen sie sich mit ... (ex. de Fleischer, 1982)
← *sich den Kopf zerbrechen* (se casser la tête)
\+ *Köpfchen haben/etwas im Kopf haben* (être très intelligent)

Dans ce qui suit, nous donnons pour le français et l'allemand des exemples de règles de grammaires locales écrites en IDAREX, et ce, pour chacune des variations et des modifications décrites en section 2.

4.1. Variations lexicales

Différents items lexicaux, généralement sémantiquement équivalents, peuvent être autorisés :

F : *perdre la tête/la boule/les pédales ≠ perdre la tronche*
⇒ perdre Verb: [:la :tête | :la :boule | :les :pédales]
A : *eine ruhige/sichere Hand (haben)* ((avoir) la main sûre)
≠ *eine stille Hand (haben)* ((avoir) la main tranquille)
⇒ :eine [:ruhige | :sichere] :Hand

On peut exprimer de façon analogue le fait qu'un argument interne à l'expression idiomatique admette plusieurs réalisations syntaxiques. Dans l'exemple allemand suivant, aussi bien un objet prépositionnel qu'un objet accusatif peuvent être utilisés :

A : *mit den Achseln/die Achseln zucken* (hausser les épaules)
⇒ [[:mit :den :Achseln | :die :Achseln] (ZU) zucken V: | ...]

4.2. Variations morpho-syntaxiques

La variation en nombre pour le nom peut être contrôlée par une variable morphologique (par exemple *Nsg*) associée à un nombre particulier.

F : *la politique de l'autruche ≠ la politique des autruches*
⇒ :la :politique :de :l' :autruche Nsg:
A : *grüne Welle* (l'onde verte) ≠ *grüne Wellen* (les ondes vertes)
⇒ grün A: Welle Nsg:

Dans d'autres cas la variation en nombre est autorisée, comme dans :

F : *comme un coq en pâte*
⇒ :comme un Det: coq N: :en :pâte
A : *verkrachte Existenz(en)*
⇒ verkracht A: Existenz N: (perdant (personne))

Le même phénomène se produit avec d'autres catégories syntaxiques : par exemple, l'emploi du comparatif ou du superlatif va parfois rompre le sens idiomatique, comme c'est le cas dans :

F : *faire table rase ≠ faire table plus rase*
⇒ faire Verb: ADV* :table :rase
A : *reinen Tisch machen* (faire table rase) ≠ *reinsten Tisch machen* (faire table plus rase)
⇒ [machen Vfin: (ADV* NPnom) ADV* rein Apos: :Tisch
| rein Apos: :Tisch (ZU) machen V:]
A : *jds. bessere Hälfte* (son meilleur côté) ≠ *gute/beste Hälfte* (bon/meilleur côté)
⇒ POSS gut Acomp: Hälfte N:

Alors que dans d'autre cas le sens idiomatique est conservé, et une variable morphologique moins restrictive peut alors être utilisée (par exemple `A`) :

F : *ses bons/meilleurs côtés*
⇒ POSS bon A: côté N:
A : *schlechter Scherz* (mauvaise plaisanterie) / *der schlechteste Scherz*
⇒ schlecht A: Scherz Nsg:

À un niveau plus syntaxique, certains LMMs admettent n'importe quel **déterminant** (`DET`) alors que d'autres sont plus restrictifs (p. ex. `INDEFDET`).

F : *engueuler quelqu'un comme du/un/des poisson(s) pourri(s)*
⇒ engueuler V: NP :comme INDEFDET poisson N: pourri A:
A : *von einer/der Idee durchdrungen* ((être) omnubilé par une idée)
⇒ :von DET :Idee ADV* durchdrungen A:

La plupart des LMMs verbaux permettent des **variations de personnes** ; pourtant une telle variation est interdite dans certaines expressions prédicatives figées. C'est le cas dans les exemples suivants où seule la troisième personne est autorisée :

F : *les bons comptes font les bons amis*
⇒ :les :bons :comptes faire Vpl3: :les :bons :amis
A *jdm fällt ein Stein vom Herzen* (se sentir soulagé d'un poids)
≠ *jdm fallen Steine vom Herzen* (des pierres tombent du cœur de quelqu'un)
⇒ [fallen Vsg3: (ADV* NPdat) ADV* :ein :Stein :vom .Herzen | (NPdat) ADV* :ein :Stein :vom :Herzen fallen Vsg3:]

Il arrive parfois que l'on ne puisse pas avoir de **variations temporelles**[6]. Dans ce cas, seul le niveau de surface est utilisé.

F : *qui a bu boira*
⇒ :qui :a :bu :boira

6 Cette restriction semble ne concerner que les proverbes

A : *Wasser hat keine Balken* (personne ne peut marcher sur l'eau)
≠ *Wasser hatte keine Balken* (l'eau n'a pas de poutre)
⇒ :Wasser :hat :keine :Balken

4.3. Modifications

Il est souvent possible d'insérer toute une classe syntaxique de mot. Par exemple, **l'insertion d'adverbes** (ADV) est un phénomène courant :

F : *prendre souvent le taureau par les cornes*
⇒ prendre Verb: ADV* :le :taureau :par :les :cornes
A : *Sie spitzt (plötzlich) die Ohren* ((soudain,) elle tend l'oreille)
⇒ [spitzen Vfin: (ADV* NPnom) ADV* :die :Ohren | ...][7]
(besonders) hohes Tier (un (très) grand manitou)
⇒ (ADV) hoch A: Tier Nsg:

Certains LMMs permettent la **modification adjectivale** (ADJ) comme c'est le cas dans :

F : *faire un () crochet*
⇒ faire Verb: :un ADJ :crochet
A : *seine (neugierige) Nase in etwas stecken* (mettre son () nez dans quelque chose)
⇒ [POSS (:neugierige) :Nase ADV* :in NPakk (ZU) stecken V: | ...]

Certains LMMs conservent leur sens idiomatique même sous **l'opération de négation**. Dans les exemples qui suivent ceci est assuré par la variable ADV. En allemand, la négation attributive est également possible et peut être décrite comme un cas de variation lexicale du déterminant.

F : *il ne prend jamais le taureau par les cornes*
⇒ prendre Verb: ADV* :le :taureau :par :les :cornes
A : *(nicht) das Handtuch werfen* (ne pas jeter l'éponge)
⇒ [werfen Vfin: (ADV* NPNom) ADV* :das :Handtuch | ...]
(k)eine (dicke) Lippe riskieren (risquer de dire des choses qu'il ne faut pas dire)
⇒ [[:eine | :keine] (:dicke) :Lippe ADV* (ZU) riskieren V: | ...]

4.4. Variations structurelles

La **passivisation** est prise en compte par l'ordre des mots d'un V-final et la disjonction de différents composants ordonnés différemment.

F : *crever l'abcès*
⇒ [crever Verb: ADV* :l' :abcès | :l' :abcès [avoir Vsg3: :été ADV* | être Vsg3: ADV*] :crevé]]

7. La deuxième partie de l'expression régulière qui prend en compte l'ordre des mots V1/V2 n'est pas mentionnée ici

A : *endlich wurde reiner Tisch gemacht* (finalement, table rase a été faite)
⇒ [rein Apos: :Tisch (ZU) machen V: | ...]

La **topicalisation**, le **scrambling** et l'**ordre des mots pour V1/V2/V finaux** sont tous pris en compte par la disjonction de différents composants ordonnés différemment.

F : *chercher midi à quatorze heures → Midi, il ne le cherchait pas à quatorze heures*
⇒ [chercher Verb: ADV* :midi :à :quatorze :heures | :midi NP ADV :le chercher Vsg3: :à :quatorze :heures]

A : *den Vogel abschießen* (surpasser tout le monde) → *Den Vogel dabei hat dann Jan abgeschossen*
⇒ [:den :Vogel ([:dabei | :bei NPdat]) schießen Vfin: ADV* NPnom ADV* :ab Pref2: | :den :Vogel ([:dabei | :bei NPdat]) Vaux (ADV* NPnom) ADV* abschießen V: | ...]

A : *für etw. den Kopf hinhalten / den Kopf für etw. hinhalten* (payer pour quelqu'un d'autre)
⇒ [DEFPOSS :Kopf ADV* [:für NPakk | :dafür] ADV* hinhalten V: | [:für NPakk | :dafür] ADV* DEFPOSS :Kopf ADV* hinhalten V: | ...]

Le patron général qui permet de rendre compte de la variation de l'ordre des mots en allemand est le suivant :

V1/V2 : _verbe_ Vfin: (ADV* NPnom) ADV* (_libre_comps_) _figé_
V-final : _figé_ ADV* (_libre_comps_) ADV* _verbe_ V:

où `libre_comps` représente tout complément externe à l'expression idiomatique et `_figé` représente les parties fixes de l'expression idiomatique qui reste à côté du `_verbe_`.

Les variations structurelles ne sont, pour l'instant, pas encodées pour le français alors que certaines d'entre elles le sont pour l'allemand. Cependant elles soulèvent un point intéressant quant aux règles de grammaire locales et au formalisme des états finis. C'est ce sur quoi nous nous penchons maintenant.

4.5. Pouvoir d'expression des règles de grammaire locales

4.5.1. L'ordre des mots

La plupart des variations et des modifications permettant de distinguer les LMMs de séquences totalement figées s'expriment aisément et naturellement à l'aide des règles de grammaire locales. Cependant, rendre compte de la variation de l'ordre des mots et des phénomènes qui y sont liés, comme la topicalisation et le scrambling, est une tâche plus délicate. Or tous ces phénomènes sont courants en allemand et il est donc primordial de pouvoir en fournir la description la plus appropriée possible. Ici, le pouvoir d'expression des règles de grammaire locales semble atteindre ses limites dans la mesure où il ne s'agit plus de décrire des phénomènes « locaux ».

Pour décrire tous les LMMs verbaux allemands avec IDAREX et dans le cadre de la technologie des états finis en général, il faut rendre compte de tous les agencements possibles des constituants, y compris toutes les positions où (modification externe) des adverbes peuvent être insérés. Une telle description peut être longue et laborieuse pour les verbes qui admettent des compléments au datif et à l'accusatif, et ce plus particulièrement, lorsque topicalisation et scrambling sont autorisés. Non seulement de telles expressions sont pénibles à lire et à décrire, mais en plus la compilation des réseaux qui leur sont associés est longue. D'autre part, la définition de variables pour les éléments pouvant être insérés nécessite une description partielle de la syntaxe de l'allemand, tout particulièrement pour ce qui concerne les constructions GN et GP.

Une alternative pragmatique et moins coûteuse du point de vue informatique consiste à regrouper dans une même variable ANY tous les constituants pouvant apparaître entre le verbe et ses compléments figés, y compris les arguments externes non idiomatiques. Bien que, dans certains cas, cette approche conduise à l'identification erronée de patrons, elle est assez fiable pour avoir été utilisée avec succès dans le cadre de COMPASS.

4.5.2. L'accord

Pour certains LMMs il est nécessaire de contrôler l'accord entre les différents constituants. Ainsi l'expression française *casser sa pipe* perd son sens idiomatique si le réflexif n'est pas à la même personne que le sujet. Un exemple analogue est donné par l'expression allemande *seine Meinung ändern* (changer d'avis). L'accord est marqué explicitement dans les règles (variable particulière) bien que l'implantation n'en fasse, pour le moment, pas l'usage.

Les deux phénomènes décrits précédemment, l'ordre des mots et l'accord, n'appartiennent plus à la classe des phénomènes locaux. Idéalement un mécanisme plus puissant serait nécessaire pour décrire leur régularité d'une façon générale. Pourtant, notre approche a été en mesure de couvrir au moins les cas les moins complexes.

5. Les applications d'IDAREX

Identifier les LMMs est essentiel pour tout traitement du langage naturel basé sur des informations lexicales : ceci est vrai pour des applications comme la concordance ou l'indexation intelligente, la traduction automatique, ou la consultation automatique de dictionnaires. Dans cet article, nous avons montré comment leur description peut être faite à l'aide d'IDAREX, en construisant des règles de grammaire locales exprimées sous la forme d'expressions régulières dans le formalisme des états finis.

Les règles de grammaires locales ici décrites sont utilisée dans LOCOLEX[8], l'outil automatique d'aide à la compréhension développé chez Rank Xerox, et utilisé

8. Pour une description plus complète de LOCOLEX le lecteur pourra consulter Bauer, Segond et Zaenem (1995).

dans le cadre du projet COMPASS. Son principal propos est de fournir à l'utilisateur une consultation automatique en contexte d'un dictionnaire de compréhension bilingue. Imaginons, par exemple, un anglophone ayant une certaine connaissance de l'allemand qui lit un texte électronique en allemand et à qui il manque certains mots du texte. Lorsqu'il clique sur un mot inconnu COMPASS renvoie non pas à l'entrée du dictionnaire dans son entier mais uniquement à la partie de l'entrée pouvant aider à la compréhension du mot dans ce contexte bien précis. Par exemple, COMPASS donnera uniquement les traductions associées à la partie du discours appropriée ou, dans le cas d'un LMM, la traduction associée à l'expression. Pour réaliser cela nous avons enrichi des dictionnaires électroniques[9] avec des règles de grammaire locales.

Un certain nombre d'applications existantes peuvent être améliorées grâce à l'utilisation de règles de grammaire locales. Ainsi les grammaires des dates peuvent améliorer les performances des systèmes optiques de reconnaissance des caractères.

Plus généralement, les règles de grammaire locales sont utiles à l'analyse syntaxique, p. ex. la description d'expressions adverbiales complexes telles que les dates en français (*le lundi 21 août au matin*)[10] ou toute expression n'obéissant pas à la syntaxe générale. Nombreux sont les cas où l'analyseur syntaxique échouera tout simplement parce qu'incapable d'analyser correctement le LMM inclus dans une phrase plus large. Par exemple, en allemand, la syntaxe générale demande qu'un nom quantifiable soit précédé d'un déterminant. Cette règle est enfreinte dans le LMM *von Haus aus* (originale).

La prochaine étape pour l'amélioration de l'analyse syntaxique consiste à incorporer les règles de grammaire locales dans un composant syntaxique général.

En ce qui concerne la technique que nous utilisons, le formalisme à deux niveaux dans un système à états finis ainsi qu'IDAREX, elle a l'avantage de fournir une représentation compacte. Comme nous l'avons vu elle donne la possibilité de définir des variables générales telles que « n'importe quel adverbe » (ADV) ou des variables morphologiques plus spécifiques telles que « seulement la troisième personne singulier du verbe » (Vsg3). Cela soulage le lexicographe de la pénible tâche d'expliciter toutes les formes possibles. Les variables fonctionnelles, ou macros, permettent d'exprimer des généralisations pour des patrons qui sont attachés à toute une classe de mots. D'autre part, les deux niveaux nous permettent d'exprimer des faits soit sur la forme de surface, soit sur la forme lexicale. Ainsi lorsque l'on veut exprimer qu'une forme est figée on n'utilise que le niveau de surface, évitant par là même, de se perdre dans tous les traits du niveau lexical.

Cette technologie permet d'effectuer des opérations sur les réseaux engendrés par les expressions régulières : addition, soustraction, intersection et composition. Bien que nous n'ayons pas encore utilisé cette possibilité dans notre travail, elle donne un grand pouvoir d'expression. Imaginons, par exemple, que nous souhaitions

9 Dans ce projet nous avons utilisé les dictionnaires suivants : le *Dictionnaire anglais-français Oxford-Hachette* (1994) et le *Dictionnaire allemand-anglais Harper-Collins* (1991). Nous sommes particulièrement reconnaissantes aux éditeurs qui ont accepté de nous fournir ces dictionnaires à des fins de recherche.

10 Un traitement analogue pour les adverbes de date en français est proposé par Denis Maurel (1993).

prendre en compte la sémantique des LMMs et pour ce faire nous ayons besoin de restreindre les règles écrites. Une possibilité qui s'offre à nous consiste à écrire de nouvelles expressions régulières et de soustraire les réseaux ainsi générés à ceux construits précédemment. De telles expressions régulières pourraient ainsi décrire la compatibilité entre les classes sémantiques des noms et des adjectifs.

Lexicographie bilingue informatisée au quotidien : témoignage du rédacteur face à l'écran

Thomas SZENDE et Dominique RADANYI

INALCO, Paris et CIEH, Université de Paris III, France

• *Abstract* •

The only Hungarian-French dictionary in use today is obsolete, i.e. *its "hungarocentrism" is obvious ; it is full of archaic meanings, lacks semantic indicators, and so on. It seemed essential to create a new dictionary both for Hungarian and French language communities, taking into account their particular needs, and to give every possible information – semantic, grammatical, stylistic and even cultural – on the lexical units and their considerations in discourse. Compiling this dictionary required the organization of a technical structure allowing it not to be a static product but a tool open to reactualization and research procedures. This paper introduces the computer tools used during the various steps of the compiling work and which will eventually be used in the electronical version of the dictionary. 1) data collecting ; headword listing (WORD-CRUNCHER) ; 2) article editing (WRITER STATION) ; 3) dictionary database (PAT).*

Nous allons présenter quelques aspects informatiques et pratiques de la préparation d'un nouveau dictionnaire hongrois-français[1].

À l'initiative du professeur *Jean Perrot*, directeur du Centre Interuniversitaire d'Études Hongroises à l'Université de Paris III, deux équipes lexicographiques ont été constituées : une en Hongrie à l'Université de Szeged, sous la direction de *Miklós Pálfy* réalisant la partie français-hongrois et l'autre à Paris, au sein du CIEH pour la partie hongrois-français sous la direction *de Thomas Szende*. L'édition de l'ouvrage serait confiée aux *Éditions Akadémiai Kiadó* (Budapest) et aux *Éditions le Robert* (Paris).

Les seuls ouvrages de référence qui existent actuellement, le dictionnaire d'*Au-*

1. Nous tenons à remercier tous nos collaborateurs (*Joëlle Dufeuilly, Viktória Eröss, Károly Ginter, Emilie Molnos, Jean-Léon Muller et Péter Zimonyi*), et plus particulièrement *Chantal Philippe* qui outre ses activités de lexicographe prend en charge la gestion informatique des fichiers au sein de l'équipe parisienne

rélien Sauvageot, publié dans les années trente et celui de *Sándor Eckhardt*, publié dans les années cinquante, sont largement dépassés pour des raisons évidentes. Les dernières générations de hungarophones étudiant le français et de francophones étudiant le hongrois sont largement tributaires de ces deux ouvrages : nous leur devons de bonnes observations sur les deux lexiques mais aussi des erreurs de traduction qu'il faut corriger et des absences douloureuses qu'il faut combler.

Dès le début des travaux, il nous est apparu fondamental qu'un nouveau dictionnaire :

- reflète fidèlement l'état actuel des deux langues dans leurs différents registres ;
- soit conçu en fonction des besoins des deux communautés linguistiques, hungarophone et francophone ;
- donne un maximum de renseignements sémantiques, grammaticaux, stylistiques et même culturels sur les unités lexicales retenues et leur fonctionnement dans le discours.

De plus, le contraste entre une langue comme le français et le hongrois, langue agglutinante, appartenant à la famille des langues finno-ougriennes, renforçait pour nous le besoin d'un dictionnaire nouveau, réalisé à partir de cette approche lexicographique particulière.

Un dictionnaire bilingue n'est pas la description lexicale exhaustive de deux états de langue, ni la reproduction fidèle des innombrables réalisations concrètes des discours en langue source et en langue cible. Il a pour fonction de mettre en parallèle les lexiques des deux langues, en apportant à l'utilisateur, à travers un nombre limité d'exemples pertinents, le moyen de produire des énoncés les plus naturels possible et d'éviter au maximum des erreurs d'interprétation.

À cette fin, l'ouvrage envisagé devra enregistrer un vocabulaire général d'environ 40 000 à 50 000 mots ; il sera complété par des lexiques spécialisés. Il va de soi que la nomenclature de ce dictionnaire contiendra obligatoirement un noyau de mots usuels, pleinement représentatif du fonds culturel, mais aussi de très nombreux termes illustrant la richesse et la diversité des langues actuelles constamment nourries des apports nouveaux de la civilisation.

Ancrée, engluée dans de longues traditions, la lexicographie se renouvelle aujourd'hui grâce à l'ordinateur[2]. Elle absorbe l'outil informatique à plusieurs niveaux ; aussi bien au niveau de l'analyse préliminaire du matériel de référence qu'à celui de la constitution du texte dictionnairique proprement dit et de la production de textes imprimés par la composition automatisée et programmée.

2 P Imbs l'a déjà constaté il y a un quart de siècle « Dans l'état actuel du monde, il nous a semblé que la machine était encore serve du livre et non pas son substitut elle déleste l'homme de tâches serviles, notamment dans le domaine de la documentation, qu'elle aide à maîtriser lorsqu'elle est. . surabondante et constamment foisonnante , elle peut aussi l'aider à poser et à résoudre des problèmes de nature quantitative ou même qualitative, en lui fournissant p ex des listages qui, pour grossière que soit leur approche, n'en facilitent pas moins les analyses fines, qui sont l'essence même de la science exacte » *Trésor de la Langue Française*, CNRS, 1971, p XIII.

La mise à profit de l'informatique suppose une analyse rigoureuse et exhaustive de la démarche méthodique du lexicographe, analyse qui non seulement tient compte mais encore explique la totalité des décisions prises par le lexicographe au cours de sa production[3]. L'informatique permet d'enregistrer l'ensemble des articles du dictionnaire, de les stocker dans une base de données, de prévoir et d'harmoniser toute intervention du lexicographe.

Les différentes structures d'articles, ainsi que les variantes probables doivent être identifiées, classifiées et répertoriées ; toutes les sections du dictionnaire sont ainsi « étiquetées ». Ce travail doit aboutir à la rédaction d'une grammaire formelle décrivant la structure du dictionnaire avec des codes identificateurs de données à chaque changement typographique, ce qui donne une structuration arborescente des articles.

En standardisant les différents champs constitutifs de chaque type d'article, il devient possible de faire appel à des procédures de recherche et d'interrogation.

L'idée même d'aide informatique à la rédaction d'un dictionnaire bilingue repose sur le principe d'une décomposition de l'ensemble du processus lexicographique en opérations et en parcelles d'opérations.

Autrement dit, on précise non seulement le contenu du dictionnaire mais aussi, et parallèlement, la structure à donner à ce contenu.

L'élaboration de notre dictionnaire par les deux équipes a nécessité la mise en place d'une structure technique et informatique permettant :

- d'une part, que le manuscrit en gestation et, plus tard, l'ouvrage soient l'objet d'une mise à jour permanente et,
- d'autre part, qu'une version électronique puisse être envisagée.

L'ensemble de programmation doit permettre de prendre en charge les exigences spécifiques d'un dictionnaire bilingue, susceptible d'assurer la cohérence et l'homogénéité de la nomenclature et d'automatiser au mieux chacune des démarches que comporte la rédaction d'un tel ouvrage.

À cette fin, nous exploitons quotidiennement deux outils informatiques conçus pour des PC (retenus après de nombreux essais infructueux avec différents logiciels).

- Pour la collecte de données et l'établissement de la nomenclature : *WORDCRUNCHER* qui permet :

  - la recherche des éléments en contexte dans le corpus (ce contexte peut se réduire à quelques lignes ou être élargi à une page-écran dont la ligne centrale contient l'occurrence recherchée) ;

3. « L'ordinateur étant idiot, on ne peut pas s'en servir sans tout lui expliquer, ce qui suppose un immense travail métalexicographique jamais encore fourni.» F J Hausmann, « La métalexicographie à l'échelle mondiale », *Coloquio de Lexicografía*, Verba, Anexo 29, Universidade de Santiago de Compostela, 1988, pp 79-109

- de connaître d'une manière très précise les différentes nuances d'un terme ou de cerner l'emploi des termes que l'on ne trouve pas dans les dictionnaires existants ; et
- de générer automatiquement des tableaux de fréquences lexicales.
- Pour la saisie et la rédaction des articles : *WRITER STATION* sur lequel nous reviendrons ultérieurement.

Nous avons également à notre disposition le logiciel *PAT* pour toute consultation de la base de données réunissant l'ensemble des articles déjà rédigés par les deux équipes. Ce logiciel est accessible à l'Institut de Linguistique de l'Académie des Sciences de Budapest. Nous devons en effet la mise au point de cet ensemble informatique à *Júlia Pajzs*, chercheur dans cet institut.

Notre méthode, décrite sous tous ses angles dans un protocole de rédaction, se perfectionne sans arrêt à l'usage. Nous l'avons voulue à la fois stricte et souple afin qu'elle soit valable pour le plus grand nombre de cas possibles.

Les spécifications de la rédaction ne pouvaient pas être fixées *a priori*. Seul un processus itératif d'essais pouvait permettre, en partant de spécifications initiales forcément approximatives, d'aboutir à des spécifications adéquates. La rédaction de plusieurs centaines d'articles tests, entre 1991 et 1993, a mis en évidence les insuffisances de la « grammaire » auxquelles il a fallu remédier.

La grammaire en question fait appel au langage standard et généralisé de marquage qui porte le nom de SGML.

Le travail de rédaction proprement dit a commencé ainsi en septembre 1993 et continue grâce à une équipe stable composée de linguistes bilingues et de traducteurs.

Voici très rapidement la manière dont nous avons organisé le travail de rédaction.

Dans une première phase, les rédacteurs hungarophones préparent la nomenclature comportant :

- les unités lexicales destinées à apparaître dans le corps du dictionnaire comme vedettes ;
- les exemples qui devraient illustrer leur emploi et leur place dans le discours ;
- les locutions figées les plus courantes autour du mot-vedette ;
- les indications sémantiques en langue hongroise nécessaires à la distinction des sens ;
- les marques d'emploi relatives au domaine et registre d'utilisation.

Dans une deuxième phase, les rédacteurs francophones :

- sélectionnent les équivalents les plus pertinents ;
- proposent une traduction des exemples et des locutions figées ;
- ajoutent s'il y a lieu des indications sémantiques et les marques d'emploi pour la partie française ;
- suggèrent des modifications dans la structure des articles en fonction des critères sémantiques du français.

Dans une troisième phase, les rédacteurs hungarophones et francophones réexaminent ensemble chaque projet d'article et adoptent une version quasi définitive.

La connaissance de trois langues est indispensable pour la rédaction de ce dictionnaire bilingue : langue source (hongrois) + langue cible (français) + langue du logiciel.

En effet, l'exploitation de ce logiciel n'étant pas une affaire d'informaticiens, il est crucial que des non-spécialistes comme nous puissent la maîtriser sans trop de difficulté.

Le système a dû être organisé de façon à ce que le commun des mortels puisse le contrôler, quelles que soient sa langue maternelle et ses compétences en informatique.

Grâce à WRITER STATION, nous bénéficions d'un tel système, fondé sur le dialogue, dans un environnement interactif, convivial, adapté à nos besoins ponctuels.

Les yeux fixés sur le masque visible dans la partie inférieure de l'écran, le claviste, qui est toujours lexicographe en même temps, dispose d'un menu pour chaque étape de son travail rédactionnel.

Voici la structuration la plus simple du cheminement informatique : il s'agit d'un article ne comportant qu'une vedette hongroise, sa catégorie grammaticale et un seul équivalent français avec son genre grammatical :

ARTICLE/ARTICLE RENVOI
|
ENT
|
VEDETTE
\
FIN VDT
\
FIN ENT/PHON/**CATÉGORIE**/RCT/DDS/RDL/LIG
|
FIN CGR/RCT/DDS/RDL/LIG
|
FIN ENT/RCT/DDS/RDL/LIG
/
FIN ART/BLS/BGR/LOC
|
FIN BLS/GEN/BEQ/DDS/RDL/LIG/IDS/RCT/EXP
\
ÉQUIVALENT
|
FIN EQV
|
GENRE/DDS/RDL/LIG/IDS/FIN BEQ/EQV
|
FIN GEN
\
DDS/RDL/LIG/IDS/FIN BEQ/EQV
\
FIN BLS/RCT/BEQ/EXP/LOC
|
FIN ARTICLE

Guidé par WRITER STATION tout au long de la saisie de l'article, le lexicographe n'est concerné que par le haut de l'écran où s'effectue son travail.

Mais, comme la pratique lexicographique le montre, il arrive rarement qu'à un terme A de la langue source corresponde précisément un terme B de la langue cible. Étant donné cette complexité de la réalité linguistique, nous avons recours à différents types d'usages du logiciel. En voici quelques exemples présentés sous leurs deux aspects :

– non formaté avec les jalons textuels qui ouvrent et ferment les champs successifs ; et
– formaté, tels qu'ils apparaîtront dans la version papier du dictionnaire.

L'article *alátét* ne comporte qu'un seul bloc sémantique, à l'intérieur duquel des indications sémantiques distinguent les différents équivalents :

<ART><ENT><VDT>alátét </VDT><CGR>n </CGR></ENT><BLS><IDS>(íróasztal) </IDS><BEQ><EQV>sous-main </EQV><GEN>m; </GEN></BEQ><BEQ> <IDS>(csavar) </IDS><EQV>rondelle </EQV><GEN>f; </GEN></BEQ><BEQ> <IDS>(pohár) </IDS><EQV>dessous </EQV><GEN>m </GEN><EQV>de verre; </EQV> </BEQ><BEQ><IDS>(edény) </IDS><EQV>dessous-de-plat </EQV><GEN>m; </GEN></BEQ><BEQ><IDS>(étkészlet) </IDS><EQV>set </EQV><GEN>m </GEN>(de table) </BEQ></BLS></ART>

alátét *n* **1** (*íróasztal*) sous-main *m* ; (*csavar*) rondelle *f* ; (*pohár*) dessous *m* de verre ; (*edény*) dessous-de-plat *m* ; (*étkészlet*) set *m* (de table)

L'article *alatti* se divise en plusieurs blocs sémantiques ; les divisions sont justifiées par des indications sémantiques et illustrées par des exemples :

<ART><ENT><VDT>alatti </VDT><CGR>adj </CGR></ENT><BLS><IDS> (hely) </IDS><EXP> kép ~ szöveg </EXP><TRD>légende <IDS>(d'un dessin, d'une photo); </IDS></TRD><EXP>vminek a víz ~ része </EXP><TRD>la partie immergée de qc; </TRD><EXP>Béke utca 5. (szám) ~ ház </EXP><TRD>l'immeuble du/situé au 5 rue Béke; </TRD><EXP>az elsõ pont ~ rendelkezések </EXP> <TRD>les dispositions mentionnées au point un</TRD></BLS><BLS><IDS>(idõ) </IDS><EXP>óra ~ beszélgetés </EXP><TRD>bavardage pendant les cours; </TRD> <EXP>a háború ~ nélkülözések </EXP><TRD>les privations de la guerre </TRD></BLS><BLS><IDS>(szint) </IDS><EXP>tíz fok ~ hõmérséklet </EXP> <TRD>température inférieure à dix degrés; </TRD><EXP>két perc ~ eredmény </EXP> <TRD> un temps inférieur à deux minutes </TRD></BLS></ART>

alatti *adj* **1** (*hely*) **kép ~ szöveg** légende (*d'un dessin, d'une photo*) ; **vminek a víz ~ része** la partie immergée de qc; **Béke utca 5. (szám) ~ ház** l'immeuble du/situé au 5 rue Béke; **az elsõ pont ~ rendelkezések** les dispositions mentionnées au point un **2** (*idő*) **óra ~ beszélgetés** bavardage pendant les cours; **a háború ~ nélkülözések** les privations de la guerre **3** (*szint*) **tíz fok ~ hőmérséklet** température inférieure à dix degrés; **két perc ~ eredmény** un temps inférieur à deux minutes

L'article *barátkozik* est du même type mais comporte en plus des renseignements grammaticaux (rection) et stylistiques ; dans le premier bloc, nous avons indiqué par des chevrons que l'équivalent n'était qu'approximatif ; parmi les exemples on trouve également une locution figée précédée d'un marquage spécifique, le dièse :

<ART><ENT><VDT>barátkozik </VDT><CGR>v intr </CGR></ENT><BLS><RCT> ~ vkivel </RCT><BEQ><EQV><se faire un/des ami(s)> : </EQV></BEQ><EXP> könnyen ~ </EXP><TRD>se lier facilement; </TRD><EXP>nehezen ~ </EXP> <TRD>avoir du mal à se faire des amis; </TRD><TRD>être peu liant; </TRD> <EXP>csak lányokkal ~ </EXP><TRD>il n'a que des amies filles; </TRD><EXP> mív vészekkel ~ </EXP><TRD>fréquenter des artistes </TRD><LFG># <EXP> fy vel-fával ~ <RDL>péj </RDL></EXP><TRD>il fraye/se lie avec n'importe qui </TRD></LFG></BLS><BLS><RCT> ~ vmivel </RCT><EXP> ~ a gondolattal, hogy </EXP><TRD>se faire à l'idée que </TRD></BLS></ART>

barátkozik *v intr* **1** ~ **VKIVEL** <se faire un/des ami(s)> : **könnyen** ~ se lier facilement; **nehezen** ~ avoir du mal à se faire des amis; être peu liant; **csak lányokkal** ~ il n'a que des amies filles; **művészekkel** ~ fréquenter des artistes # **fűvel-fával** ~ ***péj*** il fraye/se lie avec n'importe qui **2** ~ **VMIVEL** ~ **a gondolattal, hogy** se faire à l'idée que

Les deux articles *bár* présentent un cas d'homonymie. D'autre part, l'un des deux comprend plusieurs blocs grammaticaux :

<ART><ENT><VDT>1 bár </VDT><CGR>n </CGR></ENT><BLS><BEQ><EQV>bar </EQV><GEN>m; </GEN><EQV>boîte </EQV><GEN>f </GEN><EQV>de nuit </EQV></BEQ></BLS></ART>

1 bár *n* **1** bar *m*; boîte *f* de nuit

<ART><ENT><VDT>2 bár </VDT></ENT><BGR><CGR>conj </CGR><BLS> <BEQ><EQV>bien que +subj; </EQV><EQV>quoique +subj; </EQV><EQV>encore que +subj; </EQV><EQV>alors que : </EQV></BEQ><EXP> ~ nem szép, mégis sokan udvarolnak neki </EXP><TRD>bien qu'elle ne soit pas (vraiment) belle/une beauté, elle est très courtisée; </TRD><EXP>segítek rajta, ~ nem érdemli </EXP> <TRD>je l'aide, quoiqu'il ne le mérite pas; </TRD><EXP> ~ részletes, mégsem teljes a felsorolás </EXP><TRD>bien que détaillée, la liste est incomplète </TRD> </BLS></BGR><BGR><CGR>adv </CGR><BLS><BEQ><EQV>pourvu que +subj; </EQV><EQV>si seulement : </EQV></BEQ><EXP> ~ így lenne ! </EXP> <TRD>pourvu que cela se passe ainsi !; </TRD><EXP> ~ ne tette volna ! </EXP><TRD>si seulement il n'avait pas fait cela ! </TRD></BLS></BGR></ART>

2 bár I *conj* **1** bien que +subj; quoique +subj; encore que +subj; alors que : ~ **nem szép, mégis sokan udvarolnak neki** bien qu'elle ne soit pas (vraiment) belle/une beauté, elle est très courtisée; **segítek rajta,** ~ **nem érdemli** je l'aide, quoiqu'il ne le mérite pas; ~ **részletes, mégsem teljes a felsorolás** bien que détaillée, la liste est incomplète **II** *adv* **1** pourvu que +subj; si seulement : ~ **így lenne !** pourvu que cela se passe ainsi !; ~ **ne tette volna !** si seulement il n'avait pas fait cela !

À l'instar du phénomène appelé *clôture lexicale* (Kittredge, 1983) – qui signifie grosso modo que le nombre de nouveaux termes rencontrés dans une nouvelle page

diminue rapidement et tend vers zéro ou une valeur très faible quand les pages augmentent –, on peut parler d'une *clôture rédactionnelle* : au fur et à mesure que le temps passe, se profilent des préférences lexicales, grammaticales, conceptuelles courantes dans la mise en rapport des deux langues.

Le rédacteur prend conscience de sa marge de manœuvre à l'intérieur du système.

Ainsi, notre pratique quotidienne de WRITER STATION révèle différentes fonctions à l'usage.

- Une première fonction pourrait s'appeler *traitement de texte amélioré*. Le logiciel nous permet tout rajout, reformulation et correction, et améliore les conditions et la vitesse d'exécution des opérations.
- Une deuxième fonction concerne *l'optimisation des manipulations* les plus diverses : à chaque étape, la situation du lexicographe est définie. On obtient ainsi des gains de qualité et de productivité considérables.
- Une troisième fonction, *l'aide documentaire* permet l'homogénéité des exemples et des traductions, notamment par le biais de procédures, telles que : consultation, recherches, analyse et extraction d'informations.
- Une quatrième fonction, *le contrôle*, intervient en cas de dérapage ou d'oubli : un clignotement en jaune dans la fenêtre texte rappelle l'utilisateur à l'ordre et permet ainsi une maîtrise permanente de la rédaction : le logiciel reconnaît facilement qu'un article est non conforme et les erreurs de manipulation n'empêchent pas le déroulement normal de la procédure rédactionnelle.
- Une cinquième et dernière fonction, *la gestion*, autorise l'appréciation des volumes de textes traités et la constitution rationnelle des fichiers (Gouadec, 1994 : 59-74).

Une telle saisie des articles aboutit à un *système clos et ouvert* à la fois :

- le système est clos dans la mesure où tout est répertorié : on prend conscience des éventuels écarts ou informations déficitaires, et on est appelé à combler les manques ;
- le système est ouvert car l'informatisation permet toutes sortes de réorganisations cohérentes des informations données et autorise à tout moment des radioscopies insolites.

Une autre spécificité de la saisie informatique des articles est ce qu'on pourrait nommer *appel de l'équivalent* : une fois le champ ouvert, le lexicographe est tenu d'obéir au système et de compléter le canevas préétabli en fournissant une traduction, même si rien ne garantit que l'emploi qui sera fait dans les textes des mots enregistrés sera limité aux équivalents et exemples présentés.

Ceci implique une constante nécessité de réagir. Dans l'idéal, une traduction devrait rendre le sens exact, mais aussi la connotation, telle ou telle allusion ou référence culturelle, ou tels effets pouvant se situer au niveau du signifiant, comme des allitérations, par exemple. À supposer que, pris isolément, chacun de ces aspects est traduisible, tout n'est cependant pas transmissible. En principe, il est exclu qu'il y ait dans les ressources de la langue cible une équivalence où se retrouvent justement tous les aspects qui coïncident dans l'unité lexicale de la langue source.

De cette façon, nous espérons réduire au minimum les échecs de consultation, les faux-sens, les contresens et les recherches infructueuses.

L'ordinateur est un dispositif qui aide la rédaction, mais l'activité traduisante du lexicographe demeure traditionnelle : il oublie les signifiants de la langue source, tout en retenant les éléments de signification pour pouvoir les faire réapparaître grâce à des signifiants nouveaux dans la langue cible. La machine se contente d'assister la traduction humaine. C'est toujours le lexicographe qui sélectionne, traduit, agence, même si les sources à sa disposition sont plus que jamais multiples. Pondérer, filtrer, nuancer, il n'y a que le linguiste et le traducteur qui puissent le faire.

D'autre part, il nous a été impossible de prévoir et de faire apparaître tous les aspects de l'organisation interne des articles. Le masque est figé, alors que la langue est extensible.

Un des enseignements principaux que l'on tire de la pratique lexicographique est que les langues naturelles sont d'une texture à défier les formalisations et les systématisations les plus ingénieuses. En effet, les langues naturelles sont faites :

a) d'analogies et d'anomalies ;
b) de polymorphies et de polysémies ;
c) de redondances et de déficiences ;
d) d'explicitations et d'implications ;
e) de constantes et de variantes.

Le rédacteur doit donc compenser en permanence les défaillances du système. Il intervient à la fois en amont (il faut s'assurer que les articles qu'il crée soient traitables par la machine) et en aval : il faut rectifier les erreurs de l'ordinateur. Il y a nécessairement des formes « interdites » que la machine considère comme une anomalie. Il est donc nécessaire d'adapter les articles aux contraintes du logiciel.

Sans vouloir être en totale contradiction avec l'esprit même de ce colloque, nous devons avouer que la rédaction d'un dictionnaire bilingue reste un travail éminemment artisanal.

On dit souvent que la situation de tout traducteur est, par essence, exceptionnelle. De la même façon, nous avons observé que chaque cas lexicographique est unique. Les informaticiens ont résolu de nombreux problèmes posés par la technique lexicographique, mais pas les problèmes posés par les langues elles-mêmes.

Cependant, l'informatique a un impact extrêmement fort sur la réflexion méthodologique. Grâce à elle, notre discours sur le dictionnaire et notre pratique de la lexicographie bilingue sont devenus algorithmiques.

Abréviations :

ART = ARTICLE
ENT = ENTRÉE
BEQ = BLOC ÉQUIVALENT

BGR = BLOC GRAMMATICAL
BLS = BLOC SÉMANTIQUE
CGR = CATÉGORIE GRAMMATICALE
DDS = DOMAINE DE SPÉCIALITÉ
EQV = ÉQUIVALENT
EXP = EXEMPLE
GEN = GENRE
IDS = INDICATION SÉMANTIQUE
LFG = LOCUTION FIGÉE
LIG = LIMITATION GÉOGRAPHIQUE
LOC = LOCUTION FIGÉE
PHON = TRANSCRIPTION PHONÉTIQUE
RCT = RECTION
RDL = REGISTRE DE LANGUE
TRD = TRADUCTION
VDT = VEDETTE

Orientation de combinants dans les langues de spécialité : comparaison entre l'anglais et le français

Patricia Thomas et Frank Knowles

Terminologue indépendante, Cranleigh et Aston University, Birmingham, Grande-Bretagne

Introduction

Cette analyse vise à établir ce que nous nommons l'*orientation* d'une collocation. Le terme *orientation* est pertinent pour deux raisons : premièrement, du point de vue dictionnairique, le terminologue doit indiquer en entrée une partie de la collocation et, deuxièmement, l'entrée est, pour le terminologue, le principal point de référence lors de la collecte des données. Les groupes d'utilisateurs doivent pouvoir sélectionner les données selon leurs besoins. Étant donné le nombre fini d'unités lexicales qui composent une phrase, c'est donc la deuxième tâche qui est plus difficile à cause du nombre infini d'utilisateurs (même de groupes d'utilisateurs).

1. L'orientation – comment la définir et l'identifier ?

L'orientation a pour but de trouver l'entrée d'une phrase afin que le terminologue puisse l'inclure dans le dictionnaire. Ceci permet au chercheur, ou au lecteur, de retrouver la phrase dès la première recherche. Un tel choix est basé sur le **contexte** de la phrase ; en ce qui concerne la construction **verbe + proposition**, c'est ou le verbe ou la proposition qui peut être choisi comme entrée. Si la proposition est un terme composé, il faut identifier d'abord l'orientation de celui-ci selon le contexte, qui dépend à son tour des besoins de l'utilisateur. La tâche nécessite alors deux analyses.

2. Le combinant

La collocation que nous avons examinée à fond est celle du **verbe + substantif**. Cette construction est certes fort utile aux traducteurs et aux chercheurs écrivant des articles dans une langue étrangère.

La construction verbe + substantif a été nommée *combinant* en langues de spécialité. Le terme *combinant* permet une plus grande variation que le terme *collocation* du point de vue de la syntaxe (par exemple, certains adjectifs et adverbes font partie intégrale d'une phrase, tandis que d'autres, soi-disant « libres », ne le font pas, p. ex. *the vaccine was genetically-engineered* mais *the woman was attractively dressed*).

3. Les corpora

L'étude se fonde sur l'analyse de deux corpora de sciences biologiques, l'un en français et l'autre en anglais, constitués de livres didactiques et de comptes rendus de colloques sur la virologie et la bactériologie, émanant de grandes maisons d'édition et d'organisations internationales telles que l'OMS et l'OCDE. Chacun des corpora contient un demi-million de mots. Les domaines traités sont donc très restreints et les textes ne sont pas des chimères prototypiques. Cependant, étant donné que ces domaines sont en plein essor, les textes contiennent beaucoup de néologismes. Les analyses effectuées dans les deux langues ont permis de constater, d'un point de vue statistique, les conclusions d'orientation terminologique ou dictionnairique des combinants.

4. L'analysateur de textes

L'analyse des textes a été effectuée au moyen de l'Aston Text Analyser (ATA[1]). Cet analyseur fournit des listes de fréquences de mots ainsi que des concordances qui donnent quatre mots de chaque côté du mot-clé, triés par ordre alphabétique à droite ou à gauche.

Pour l'ATA, ce qui est important, c'est le *profil synoptique* dans lequel le mot-clé est indiqué par un astérisque. Dans le tableau 1, les trois colonnes à gauche et à droite du mot-clé contiennent les mots qui apparaissent en positions -3 à -1 et +1 à +3 respectivement, en ordre décroissant de fréquence.

Du profil synoptique on peut passer directement à la section du texte dans laquelle se trouve le mot-clé ; ces quelques lignes de texte vont donc au-delà du niveau de la phrase.

5. Comparaison des verbes « de spécialité » en anglais et en français

Le pourcentage total de tous les verbes dans les corpora – et non seulement ceux qui sont représentatifs du domaine – dépasse les 8 %. Il convient de noter la rareté des verbes dits « de spécialité » dans les deux langues ; en anglais, la proportion n'est que 0,02 % du nombre total de mots. On appelle « verbes de spécialité » ceux qui n'existent que dans un domaine spécifique ou dont l'usage est significativement plus fréquent que d'ordinaire. Dans les deux corpora, se trouvent 28 verbes de spécialité anglais et 24

1 Développé par le Dr Peter Roe, le professeur Frank Knowles et leurs collègues à l'Université d'Aston de Birmingham en Grande-Bretagne, avec leur partenaire MS Technology A/SA à Copenhague au Danemark.

français en hapax. Le tableau 2 indique la fréquence des verbes de spécialité dans les sciences biologiques en anglais et en français ; ce qui est intéressant, c'est que le plus grand nombre de verbes dans leur forme non lemmatisée apparaît moins de cinq fois. En plus, en anglais on trouve quelques occurrences de verbes composés (par exemple *to phase-vary*). Ces verbes sont en général des verbes dénominalisés et, étant donné le progrès scientifique rapide et continu dans ces domaines, ils sont souvent des néologismes qui finissent par être adoptés dans la langue écrite.

6. Critères d'analyse et exemples

Une analyse a été faite des combinants, verbe + substantif, qui se présentent plus de trois fois (chiffre purement arbitraire), sans tenir compte de la position du substantif par rapport au verbe. Des substantifs peuvent se trouver en positions +1, +2, +3 et +4. La place du verbe est même parfois plus éloignée.

Des exemples sont donnés du verbe transitif *express* en anglais et du verbe transitif/réflexif (*s'*)*exprimer* en français (tableaux 3a et 3b).

7. La valence des verbes de spécialité

Nous avons incorporé la théorie de valence à notre recherche sur l'orientation de la collocation. Dans cette théorie, le verbe est considéré comme le noyau de la phrase, les substantifs et d'autres éléments sont en second lieu. Quoique la valence soit considérée principalement comme structure syntaxique, elle comprend néanmoins des restrictions sémantiques et, en plus, elle a l'avantage d'opérer dans les limites de la phrase. Il est intéressant de noter que des différences au niveau de la valence peuvent se trouver entre les verbes de spécialité et les verbes en langue générale, p. ex. *l'agrégat cristallise* (et non pas « se cristallise ») ; *patients present with* (symptômes). Les deux exemples ont un réflexif qui n'est pas exprimé mais qui est sous-entendu.

8. L'orientation basée sur les actants et les circonstants

L'identification des actants et des circonstants en théorie de la valence peut fournir une aide importante lorsqu'on veut constater l'orientation d'un combinant. Les circonstants se présentent souvent sous forme de syntagmes prépositionnels qui peuvent être facultatifs, dont on peut faire abstraction sans que la phrase perde son sens. Tous les combinants ont des verbes avec un actant obligatoire mais des circonstants facultatifs.

Il en résulte trois catégories de combinants auxquelles il faut accorder des rôles de *base* et de *collocataire* pour arriver à l'orientation. Premièrement, si on passe de la forme active à la forme passive, c'est l'actant obligatoire en première position de valence qui est la base, et qui fournit donc l'orientation du combinant, tandis que le verbe est la partie collocataire (tableau 4a). Deuxièmement, aux cas où la nominalisation nécessite un verbe support ayant peu de valeur sémantique, l'orientation reste sur le substantif en deuxième position de valence, car c'est celui-ci qui contient la plus

grande partie de l'information de la phrase et qui est la base du combinant (tableau 4b). Troisièmement, il est possible que les verbes avec un réflexif qui peut être sous-entendu ont le verbe comme base du combinant, le sujet ou le pronom réflexif devenant les collocataires (tableau 4c).

Conclusion

Ce travail sert à fournir des critères pour l'identification des *bases* et des *collocataires* dans les combinants des langues de spécialité, afin de faciliter le travail du terminologue. En plus, il vise à établir les fondations des *dictionnaires de collocations*.

Annexe

Span -3	**Span -2**	**Span -1**	**Type**	**Span +1**	**Span +2**	**Span +3**
2 Pir-46	2 JRS4	4 and	**expresses**	8 a	2 F	2 2
2 a	2 cells	4 that	**expresses**	5 the	2 and	2 F
2 generally	2 contain	4 which	**expresses**	2 It	2 functional	2 Pheromone
2 perhaps	2 encodes	2 constitutively	**expresses**	2 pheromone	2 genes	2 Staphylococcus
2 shown	2 more	2 efficiently	**expresses**	2 protein	2 inhibitor	2 amino
2 the	2 respectively	2 importantly	**expresses**	2 recombinant	2 multi-functional	2 protein
2 vaccinia	2 single	2 pLRO49	**expresses**	1 CR2	2 stable	2 under
2 which	2 vector	2 plr	**expresses**	1 authentic	1 Thirty-three	2 urease
1 MVA	1 Raji	1 OECD	**expresses**	1 both	1 Cells	1 As
1 Recombinant	1 The	1 presumably	**expresses**	1 its	1 RPV	1 H
1 bodies	1 VV	1 simultaneously	**expresses**	0 -	1 fusion	1 HA
1 line	1 and	0 -	**expresses**	0 -	1 haemagglutinin	1 Proceedings
1 recombinant	1 developed	0 -	**expresses**	0 -	1 immunising	1 antigen
1 selection	1 recombinant	0 -	**expresses**	0	1 influenza	1 glycoprotein
1 tested	1 that	0 -	**expresses**	0 -	1 recombinant	1 hemagglutinin
1 virus	1 vaccine	0 -	**expresses**	0 -	1 sincere	1 thanks
1 was	1 virus	0 -	**expresses**	0 -	1 the	1 were

TABLEAU 1 Exemple de profil synoptique du verbe anglais *expresses* utilisant l'Aston Text Analyser (ATA)

Fréquence	Nombre total de verbes **anglais** de spécialité (toutes formes morphologiques)	Nombre total de verbes **français** de spécialité
> 200	3	0
200 - 101	3	0
100 - 81	3	2
80 - 61	9	0
60 - 41	14	8
40 - 21	28	23
20 - 16	22	8
15 - 11	19	15
10 - 6	21	18
5 - 1	140	77

TABLEAU 2 : Fréquence des verbes spécialisés dans les sciences biologiques en anglais et en français.

Corpus examples of *express* + *cell(s)*

express	*packaging cells. A major*	(+2)
express	*inefficiently in cells in*	(+3)
express	*on their cell surface*	(+3)

Corpus examples of *express* + *gene(s)*

express	*genes from other organisms*	(+1)
express	*any genes for foreign antigens*	(+2)
express	*this gene during infection*	(+2)
express	*the IFN-gene*	(+3)
express	*the reporter gene.*	(+3)
express	*heterologous protein-specifying genes*	(+3)
express	*viral and recombinant genes.*	(+4)

Corpus examples of *express* + *polysaccharide(s)*

	(reported to) express polysaccharide were examined under	(+1)
express	*polysaccharide. Microbial polysaccharides have*	(+1)
express	*polysaccharide. Microbial polysaccharides have*	(+3)
express	*two distinctly different polysaccharide phenotypes*	(+4)

Corpus examples of *express* + *protein(s)*

express	*protein F even under*	(+1)
express	*protein F normally (compare)*	(+1)
express	*either protein F or*	(+2)
express	*heterologous protein-specifying genes for*	(+2)
express	*some proteins such as*	(+2)
express	*these proteins in L.*	(+2)
express	*multiple M-like proteins;*	(+3)
express	*the gE protein, and*	(+3)
express	*a plasmid encoded protein*	(+4)

Corpus examples of *express* + *sequence(s)*

express	*retroviral sequences One reason*	(+2)
express	*such sequences when placed*	(+2)
express	*these sequences Polytropic viruses*	(+2)
express	*VL30 sequences at high*	(+2)
express	*VL30 retroviral-related sequences*	(+3)

Corpus examples of *express* + *phenotype(s)*

express	*the ropy phenotype. Commercial*	(+3)
express	*the ropy phenotype expressed*	(+3)
express	*the ropy phenotype This*	(+3)
express	*two distinctly different polysaccharide phenotypes*	(+5)

TABLEAU 3(a) : Exemples du corpus du verbe *express* + propositions qui se rencontrent plus de 3 fois. La position de la proposition suivant le verbe est entre parenthèses

Forme transitive

un virus vivant atténué	exprimant	la GP160, une protéine d'enveloppe du HIV	(+2)
la création de bibliothèques d'ADN	exprimant	les gènes qui codent pour ces protéines immunogènes	(+2)
toutes les cellules nerveuses	expriment	la béta-galactosidase	(+2)
le dosage est	exprimé en	indiquant la quantité	(-2)
concentration de ractopamine intacte	exprimée	en chlorhydrate de ractopamine	(-2)
bacillus megaterium	exprimée	dans Bacillus subtilis	(-2)
bacillus stearothermophilus	exprimée	dans Bacillus subtilis	(-2)
la quantité de diazinon,	exprimée	en g/kg.	(-1)
référence dont la dureté,	exprimée	en carbonate de calcium	(-1)
matière caséeuse peut être	exprimée	des lésions	(-4)

Forme réflexive

on va savoir comment	s'exprime	un message génétique	(+2)
les médecins	s'expriment	peu	(-1)
les symptômes gravissimes	s'expriment	tôt dans l'enfance	(-2)
le gène marqueur	s'y exprimait	rapidement chez un homme	(-2)

TABLEAU 3(b) : Exemples du corpus français du verbe « (s')exprim* » qui paraissent plus de 3 fois. Les chiffres entre parenthèses indiquent la position avant ou après le verbe selon la (non-) réflexivité
* = *joker*

Exemple du corpus *Enzymes*	**SLOT 2** **Valence** position 1 **(BASE DU COMBINANT)**	**Actant** obligatoire
were encoded	**SLOT 1** **Verbe de spécialité** **(PARTIE COLLOCATAIRE)**	
by three genes	**SLOT 3** **Valence** position 2	**Actant** obligatoire circonstant facultatif
for sugar metabolism	**SLOT 4** **Valence** position 3	**Circonstant** facultatif

TABLEAU 4a · Base et collocation d'un verbe transitif au passif.
La base est une première position de valence.

Exemple du corpus *Fowlpox virus recombinant*	**SLOT 2** **Valence** position 1	**Actant** obligatoire
confers	**SLOT 1** Verbe support **(PARTIE COLLOCATAIRE)**	
protection	**SLOT 3** **Valence** position 2 **(BASE DU COMBINANT)**	**Actant** obligatoire
in chickens	**SLOT 4** **Sous-valence** position 1 au substantif en position de valence 2	**Actant** obligatoire *ou* **Circonstant** facultatif

TABLEAU 4b : Base et collocation d'un verbe support. La base est en deuxième position de valence.

Exemple du corpus	**SLOT 2**	
Le gêne marqueur (s'y)	**Valence** position 1 **(PARTIE COLLOCATAIRE)**	**Actant** obligatoire
exprimait	**SLOT 1** **Verbe réflexif** **(BASE DU COMBINANT)**	
rapidement	**SLOT 3** **Valence** position 2	**Circonstant** facultatif
chez un homme	**SLOT 4** **Sous-valence** position 1 au substantif en position de valence 1	**Actant** obligatoire *ou* **Circonstant** facultatif

TABLEAU 4c : Base et collocation des verbes réflexifs (parfois sous-entendus).

ACABIT : une maquette d'aide à la construction automatique de banques terminologiques

Béatrice DAILLE[1]

Université de Nantes, IRIN, Nantes, France

1. Introduction

Une banque terminologique contient le vocabulaire d'un domaine technique : les termes. Ce vocabulaire technique comprend des unités lexicales simples et des unités lexicales complexes. Parmi ces unités lexicales complexes, les noms composés sont les plus nombreux. Benveniste (1966) les a baptisés « synapsies » et les a caractérisés par un certain nombre de propriétés d'ordre morphosyntaxique et sémantique. La synapsie serait, toujours d'après Benveniste, la formation de base des nomenclatures techniques.

Le travail d'élaboration d'une banque de terminologie est un travail difficile, long et qui demande à la fois des connaissances linguistiques et terminologiques. L'enjeu est donc de fournir des outils permettant d'aider à la création de ces banques, ou quand elles existent déjà, de pouvoir les valider. Il existe deux techniques principales de dépouillement terminologique : une technique structurelle fondée sur une analyse syntaxique plus ou moins poussée de l'énoncé et une technique statistique et numérique qui décèle les associations préférentielles présentes dans les corpus.

En ce qui concerne la technique structurelle, nous pouvons citer les travaux de David et Plante (1990) et Bourigault (1992). Le logiciel TERMINO présenté dans David et Plante est un système de reconnaissance des synapsies dont les fondements théoriques s'inscrivent dans la théorie X-barre. Le module de repérage des synapsies opère sur un texte non préalablement étiqueté, à l'inverse de notre travail et de celui de Bourigault. Ce module est partie intégrante du module d'analyse syntaxique et s'appuie sur une décomposition des synapsies en tête (nom) et expansion(s) (adjectif,

1. Université de Nantes - IRIN, 2, rue de la Houssinière, 44072 Nantes Cedex 03 email : Béatrice.Daille@irin.univ-nantes.fr

groupe prépositionnel ou encore nom). Ces règles de dépistage s'appuient sur une description des marques syntaxiques de frontières et des structures grammaticales admissibles en exploitant des informations morphologiques sur la catégorisation grammaticale des mots.

À la différence de TERMINO qui effectue une analyse syntaxique de la phrase, le logiciel LEXTER, présenté dans Bourigault (1992), utilise des techniques d'analyse syntaxique locale par patron de surface. Il ne s'agit plus d'implémenter une grammaire complexe des termes mais : « de s'appuyer sur des connaissances *en négatif* concernant les configurations grammaticales dont on sait qu'elles ne peuvent pas être des constituants de termes (verbe, conjonction, pronom, etc.) ». Le texte pré-étiqueté est donc découpé en syntagmes nominaux grâce au repérage de leurs frontières potentielles. Ces groupes nominaux, dits « maximaux », ainsi que les sous-groupes qui les constituent, sont des candidats termes qu'il faudra soumettre à un terminologue. Ces deux approches, malgré la différence de leur complexité d'analyse, sont structurelles et ne permettent pas d'obtenir une liste ordonnée des candidats termes.

L'autre approche est l'approche statistique. Cette approche très prisée outre-Atlantique a donné d'excellents résultats dans le domaine du traitement du langage naturel, principalement pour la reconnaissance de la parole et pour l'assignation d'étiquettes grammaticales. Dans le cadre de l'aide à la construction de dictionnaires monolingues, l'application de modèles statistiques sur des textes fournit des informations quantitatives et qualitatives sur les affinités lexicales que peuvent présenter certains mots entre eux. Par exemple, Church et Hanks (1990) sur l'anglais, Calzolari et Bindi (1990) sur l'italien se sont intéressés aux cooccurrences lexicales mises à jour par l'utilisation d'une mesure proche du concept d'« information mutuelle », le « score d'association » (*association ratio*). Smadja et McKeown (1990), à partir d'un texte étiqueté et de l'utilisation d'une mesure similaire au score d'association, recensent les cooccurrences lexicales et les expressions figées et les intègrent, après un filtrage *a posteriori*, dans un dictionnaire utilisé par un programme de génération. À ce jour, ce sont les seuls qui produisent une application pratique de ces informations lexicales extraites automatiquement d'un corpus. Dans le domaine plus spécifique de l'acquisition de terminologies, il faut mentionner les bons résultats obtenus avec le système ANA développé par Enguehard (1992). Grâce à l'exploitation d'une liste donnée *a priori* de concepts pertinents du domaine et la mise en œuvre d'heuristiques statistiques finement ajustées, ce système extrait des concepts d'un texte avec un bon taux de précision, sans effectuer d'analyse linguistique.

Le problème principal avec l'une ou l'autre de ces approches est le « bruit ». En effet, les critères morphosyntaxiques ne permettent pas véritablement de différencier groupes nominaux libres et termes, et les cooccurrences extraites grâce à des méthodes statistiques relèvent d'associations diverses. Rappelons que parmi les cooccurrences extraites par le modèle statistique de Lafon (1984) se trouvent des associations sémantiques, des associations fonctionnelles parmi lesquelles on rencontre des noms composés ou des termes, et des associations impossibles à caractériser.

ACABIT est un logiciel de dépouillement terminologique, chargé de préparer la tâche du terminologue en lui proposant une liste ordonnée de « candidats termes », c'est-à-dire des noms composés les plus représentatifs du domaine à ceux qui le sont le moins. Il utilise des méthodes statistiques qui sont tout à fait adaptées à ce genre de

tâche puisque leurs analyses de corpus de grande taille fournissent des résultats inaccessibles à un observateur humain ou à un analyseur syntaxique et permettent de recueillir des observations générales. Il guide ces modèles statistiques sur les cooccurrences que nous voulons extraire, les termes, et évite le plus possible la prise en comptes des autres types de cooccurrences. ACABIT procède en deux étapes : d'abord il filtre les séquences morphosyntaxiques qui caractérisent les « termes de base » grâce à des grammaires locales (voir section 2.2.), puis il utilise un modèle statistique pour distinguer parmi ces cooccurrences lesquelles sont le plus probablement des termes.

2. Données linguistiques

De manière à déterminer sur quelles séquences d'unités lexicales nous allons appliquer nos mesures statistiques, nous allons utiliser les résultats d'une étude linguistique sur les structures morphosyntaxiques des termes rencontrés soit dans des corpus, soit dans des banques terminologiques existantes. Cette étude va nous permettre de dégager des spécifications linguistiques pour les termes que nous utiliserons pour établir des filtres linguistiques.

2.1. Spécifications linguistiques

Les termes sont majoritairement des unités lexicales complexes de type nominal. Nous avons voulu vérifier cette affirmation en étudiant une banque terminologique multi-domaines d'environ 800 000 termes. Il nous est impossible pour des raisons de confidentialité d'invoquer le nom de cette banque, nous la nommerons donc BANQUE tout au long de cet article.

Une première étude statistique simple portant sur la longueur des termes de BANQUE montre que 85 % de ceux-ci sont de longueur > 1 et confirme donc le fait que les termes soient majoritairement des unités lexicales complexes. Il reste donc à démontrer que ces unités lexicales complexes sont effectivement majoritairement de type nominal et d'identifier les structures morphosyntaxiques les plus représentées.

Notre deuxième étude concerne donc la représentativité des structures morphosyntaxiques des termes de BANQUE. En effet, si l'on considère les termes comme une sous-classe des noms composés, ceux-ci peuvent être caractérisés par certaines propriétés morphologiques ou/et syntaxiques mises à jour dans des études sur la composition nominale (présentées dans Gross *et al.*, (1986) ; Noailly, (1990), etc.). Plus précisément, les termes peuvent être classés en fonction de leur structure morphosyntaxique, N Adj, N1 *de* N2, etc., et s'adaptent donc à la typologie plus générale des noms composés du français élaborée par Mathieu-Colas (1988). Nous avons donc décidé d'assigner sa catégorie grammaticale à chacune des unités lexicales contenues dans les termes de BANQUE. Nous avons utilisé un logiciel développé par l'équipe de recherche en TAO du CITI[2] qui utilise le *Dictionnaire morphologique du français*

2 Centei for Information Technology Innovation (CITI), 1575 Boulevard Chomedey, Laval (Québec), Canada H7V 2X2

(*DMF*) et assigne toutes les catégories grammaticales possibles d'une unité lexicale reconnue. Les mots du *DMF* provenant essentiellement des dictionnaires *Petit Robert* et *Petit Larousse*, ils appartiennent principalement à la langue courante. La proportion des mots de BANQUE inconnus du *DMF* est de l'ordre de 30% : une grande partie du vocabulaire technique n'est donc pas reconnu. Au problème des mots inconnus se rajoute celui des ambiguïtés grammaticales. Les programmes d'étiquetage automatique tel que celui de Foster (1991) ou de El-Bèze (1993) nécessitent un contexte phrastique et ne sont pas véritablement performants sur des mots ou des courtes séquences de mots isolés. Nous avons donc développé un programme de désambiguïsation à base de règles à partir des catégories grammaticales proposées par le *DMF*. Nous avons aussi inclus dans ce programme un module chargé du traitement des unités lexicales inconnues à l'intérieur d'un terme. Nous ne détaillerons pas plus ici ce programme qui fera l'objet de publications ultérieures. Après l'application de notre programme, la proportion d'unités lexicales inconnues à l'intérieur de termes de structure complexe tombe à 3%. Il nous est donc possible d'évaluer la proportion de termes composés de plusieurs unités lexicales complexes de type nominal : celle-ci est de l'ordre de 95 %.

L'étude des termes présents dans BANQUE a donc bien démontré que ceux-ci sont pour leur grande majorité des unités lexicales complexes de type nominal.

2.1.1. Les termes de base

L'étiquetage grammatical des termes de BANQUE nous permet de classer ceux-ci en fonction de leur structure morphosyntaxique. Le tableau ci-dessous (tableau 1) présente les structures des termes les plus fréquentes (à l'exception de la structure N1 N2) ainsi que leur fréquence et leur représentativité par rapport au nombre total de terme de longueur > 1 égal à 738 072 dans BANQUE.

Structures morphosyntaxiques	**Nombres**	**%**
N Adj	182 267	25
N1 Prep N2	170 710	23
N1 Prep Det N2	46 967	6
N1 N2	14 895	2
Total	**414 839**	**56**

TABLEAU 1. Structures morphosyntaxiques des termes de base de BANQUE.

De cette étude statistique des structures morphosyntaxiques des termes, il apparaît que les termes de longueur 2, où seules sont prises en compte les unités lexicales pleines tels que les noms, les adjectifs et les adverbes séparés par des blancs dans l'écriture, sont de loin les plus nombreux. L'approche statistique exigeant une bonne représentation du nombre d'échantillons, ACABIT se concentre sur l'extraction des termes de longueur 2, appelés « termes de base », et qui s'apparient à l'une des structures morphosyntaxiques suivantes :

N Adj : *indicateur environnemental*
N1 Prep N2 : *protéine de poissons*
N1 Prep Det N2 : *chimioprophylaxie au rifampine*
N1 N2 : *bague étalon.*

2.1.2. Les termes ternaires et n-aires

Les termes de longueur > 2 sont généralement moins représentés dans les textes que les termes de base. Dans BANQUE, ils représentent pourtant environ 40 % des termes de longueur > 1. Nous pouvons néanmoins affirmer que la majorité de ces termes de longueur > 2 sont créés récursivement à partir des termes de base. Nous avons identifié deux opérations qui permettent de passer d'un terme de base à un terme de longueur > 2 : l'insertion et la juxtaposition. L'insertion, à la différence de la juxtaposition, modifie la structure morphosyntaxique du terme de base.

2.1.2.1. Insertion

Un terme obtenu par insertion est construit à partir soit d'un terme de base lorsque celui-ci est modifié, soit de deux termes de base lorsqu'il y a substitution. Dans les deux cas, la structure morphosyntaxique du terme de base est altérée.

2.1.2.1.1. Insertion de modifieurs

Ce sont principalement les adjectifs et les adverbes qui peuvent s'insérer à l'intérieur d'un terme de base ; les adjectifs dans la structure N1 Prep N2 ou N1 Prep Det N2 et les adverbes à l'intérieur d'une structure N Adj :

N1 Prep N2 → N1 **Adj** Prep N2 : *charge* ***corporelle*** *d'équilibre*
N Adj → N **Adv** Adj : *fer* ***non*** *hémique*

2.1.2.1.2. Substitution

La substitution se définit ainsi : étant donné un terme de base, l'une des unités lexicales pleines de ce terme est remplacée par un autre terme de base dont la tête est cette unité lexicale. Par exemple, dans la structure N1 Prep1 N2, N1 peut être remplacé par un terme de structure N1 Prep2 N3, pour former un surcomposé de structure N1 Prep2 N3 Prep1 N2 : par exemple, le nom *réseau* dans le terme *réseau à satellite(s)* est remplacé par le terme de base *réseau de transit* pour former le surcomposé *réseau de transit à satellite(s).*

La substitution se différencie de la juxtaposition (présentée ci-dessous) car :

– elle demande obligatoirement l'emploi de deux termes de base, dans l'exemple ci-dessus *réseau à satellite* et *réseau de transit* ;
– elle brise la structure interne de l'un des deux termes : dans notre exemple, la structure de *réseau à satellite* est altérée.

2.1.2.2. Juxtaposition

Un terme obtenu par juxtaposition est construit à partir d'un terme de base. Nous avons distingué deux sortes de juxtaposition : la surcomposition et la modification.

2.1.2.2.1. Surcomposition

La juxtaposition utilise au minimum un terme de base et se caractérise par les propriétés suivantes :

– les éléments de la structure du ou des termes de base restent solidaires ;
– lorsqu'un nom simple se juxtapose à un terme de base, c'est le plus souvent le nom simple qui précède le terme ;
– la juxtaposition s'effectue par l'intermédiaire d'une préposition ;
– les enchevêtrements à l'intérieur de la structure juxtaposée ne réfèrent pas à des termes de base. Cette propriété est illustrée dans les exemples qui suivent, où les termes de base apparaissent entre crochets :

N1 Prep1 [N2 Adj] (longueur 3)
dispositif d'[éclairage ultraviolet]
Avec *dispositif ultraviolet* qui n'est pas un terme de base.

[N1 Adj1] Prep1 [N2 Prep2 Det N3] (longueur 4)
[*accès multiple*] *avec* [*assignation à la demande*]
Ni *accès avec assignation*, ni *accès à la demande* ne sont des termes de base.

2.1.2.2.2. Postposition de modifieurs

Les termes de longueur ≥ 3 obtenus par post-modification sont les plus représentés dans BANQUE. Les adjectifs et les groupes prépositionnels adverbiaux sont les principaux modifieurs des termes unaires ou binaires qui sont à l'origine de nouveaux termes :

[N1 Adj1] **Adj2** (longueur 3) : [*production primaire*] ***épondique***
[N1 Prep N2] **Adj** (longueur 3) : [*cessation d'emploi*] ***forcée***
[N1 Adj1] [**Prep Adj N**] (longueur 4) : [*câble(s) sous-marin(s)*] [***à large bande***]

Dans un texte technique, il est difficile de déterminer si une séquence morphosyntaxique pouvant caractériser un terme de longueur > 2 obtenu par insertion ou juxtaposition est ou n'est pas un terme. Pour certaines séquences, l'introduction d'une abréviation permet d'entériner leur statut terminologique : comme par exemple l'abréviation *BLU* associée à la séquence *bande latérale unique*, obtenue par juxtaposition et plus précisément par post-modification du terme de base *bande latérale* par l'adjectif *unique* – adjectif par ailleurs très fréquent en français et qui ne porte aucune marque de technicité. Néanmoins dans la plupart des cas, il est impossible de déterminer si une séquence morphosyntaxique d'un terme de base auquel s'est appliquée l'opération de juxtaposition ou d'insertion réfère ou non à une notion du domaine. Nous avons donc décidé de ne pas trancher et de nous concentrer sur les termes de base.

Une fois ceux-ci identifiés, les « nouveaux termes » obtenus par juxtaposition pourront être reconnus à l'aide d'un programme de mise en évidence des variations terminologiques comme, par exemple, le logiciel FASTR de Jacquemin (1994).

Cependant, même en nous restreignant à l'extraction des termes de base, il nous faut prendre en compte leurs variantes.

2.1.3. Les variantes des termes de base

Il existe cinq catégories principales de variantes : les abréviations, les variantes orthographiques, les variantes morphosyntaxiques, les variantes syntaxiques et les variantes elliptiques. Les abréviations qui sont extraites des corpus par des heuristiques ne sont pas décrites ci-dessous.

2.1.3.1. Variantes orthographiques

Les variantes orthographiques d'un terme de base sont principalement de trois types :

– variation en nombre de N2 normalement invariable dans les structures N1 Prep N2 : *réseau(x) à satellite* ou *réseau(x) à satellites* ;
– l'un des composants du nom composé à plusieurs graphies possibles : ***S**ervice national* ou *service national* ;
– caractère optionnel du trait d'union dans la structure N1 N2 : *mode-paquet* ou *mode paquet.*

2.1.3.2. Variantes morphosyntaxiques

Les variantes morphosyntaxiques d'un terme de base sont principalement de trois types :

– simplification de la structure du terme binaire par l'effacement de la préposition ou/et du déterminant qui apparaît à l'intérieur de celui-ci : *tension d'hélice* = *tension hélice ;*
– relation de synonymie entre deux structures de nom composé qui diffèrent seulement par l'une de leurs unités lexicales pleines : *réseau commuté* ou *réseau à commutation* ;
– variation de la préposition : *réseau pour données* → *réseau de données.*

2.1.3.3. Variantes syntaxiques

Les variantes syntaxiques d'un terme de base sont principalement de deux types :

– insertion d'un modifieur à l'intérieur d'un terme de base : *réseau numérique* → *réseau* ***entièrement*** *numérique ;*
– coordination de deux termes de base ; *élévateurs de fréquence* + *abaisseurs de fréquence* → *élévateurs* ***et*** *abaisseurs de fréquence.*

2.1.3.4. Variantes elliptiques

Un terme peut être évoqué par une forme elliptique où une ou plusieurs de ses unités lexicales non grammaticales ont disparu. Pour les termes de base, c'est principalement l'élément de queue qui disparaît : *débit* sera employé à la place de *débit binaire*.

Cette étude linguistique a montré que, d'une part, les termes de base sont les plus représentés dans BANQUE et que, d'autre part, la majorité des termes de longueur > 2 sont construits à partir de termes de base. La décision de se concentrer sur les termes de base est donc linguistiquement motivée. À cette motivation linguistique, il faut ajouter l'argument statistique de leur bonne représentation numérique dans les corpus. Il nous reste donc à expliquer comment ACABIT extrait d'un corpus ces termes de base.

2.2. Filtres linguistiques

Nous nous trouvions devant le choix suivant : soit nous isolions, grâce aux mesures statistiques, les collocations du corpus puis nous appliquions des filtres linguistiques (à la manière du travail présenté dans Smadja et McKeown (1990)), soit nous appliquions d'abord les filtres linguistiques et ensuite les mesures statistiques. C'est cette dernière solution que nous avons adoptée. En effet, la première solution demandait l'utilisation d'une fenêtre de taille fixe ; si vous utilisez une fenêtre de petite taille, vous perdez les occurrences des termes de base modifiés par un modifieur inséré et les structures de termes de base coordonnées ; si vous prenez une fenêtre de taille plus importante, vous obtenez beaucoup de mauvaises séquences. Dans les deux cas, et même avec un filtrage linguistique postérieur, les comptes de fréquences étaient erronés. L'utilisation de filtres linguistiques avant l'application de mesures statistiques est donc la solution retenue.

ACABIT filtre les termes binaires en utilisant leurs structures morphosyntaxiques. Notre programme nécessite donc en entrée un corpus nettoyé où chaque unité lexicale a reçu son étiquette grammaticale et son étiquette morphologique (lemme). ACABIT prend en entrée uniquement jusqu'à maintenant des corpus étiquetés et lemmatisés par les programmes d'assignation d'étiquettes grammaticales et morphologiques d'IBM-France développé par l'équipe de recherche sur la reconnaissance de la parole (se référer, par exemple, aux travaux de El-Bèze (1993)).

2.2.2. Programme d'extraction et de relevé des fréquences

Les termes binaires sont considérés comme des cooccurrences particulières qui possèdent les propriétés linguistiques décrites ci-dessus : ils se définissent par rapport à leur structure morphosyntaxique ; ils ont la propriété de donner naissance à de nouveaux termes ; ils admettent des variantes. Une cooccurrence qui caractérise un terme binaire répond aux conditions suivantes :

1) elle est orientée et suit l'ordre linéaire du texte ;
2) elle met en jeu deux unités lexicales pleines ;
3) elle doit apparaître dans l'une des structures morphosyntaxiques des termes binaires.

Le relevé des fréquences des occurrences des termes de base candidats est essentiel puisque ce sont ces fréquences qui sont les paramètres des mesures statistiques. Un mauvais relevé des fréquences entraînera des résultats statistiques faux ou non pertinents pour notre application. Nous prenons en compte dans ces automates les variantes graphiques et morphosyntaxiques à l'exception des variantes elliptiques, ainsi que les variations syntaxiques qui affectent les termes de base lors des opérations de coordination et d'insertion de modifieurs. Nous avons vu en section 2.1.2. qu'un terme de longueur > 2 pouvait être obtenu à partir d'un terme de base par insertion ou juxtaposition. Si l'opération de juxtaposition ne pose pas de problème à l'extraction des termes de base puisqu'elle ne modifie pas la structure interne de celui-ci, ce n'est pas le cas de l'opération d'insertion. Une séquence comme *antenne parabolique de réception* peut référer soit à un terme de longueur 3 obtenu par insertion (soit par insertion de modifieur ou par substitution), soit à un terme de base *antenne de réception* modifié par l'adjectif *parabolique*. Si d'un côté, nous ne souhaitons pas extraire de termes de longueur > 2, d'un autre côté, nous ne pouvons pas ignorer cette variation syntaxique. ACABIT prend donc en compte l'insertion possible de modifieurs dans les structures N Adj, N1 Prep N2 et N1 Prep Det N2. L'autre variation syntaxique qui est prise en compte dans notre programme est la coordination. La coordination de deux termes de base ne produit pas, en général, un nouveau terme. Ainsi, une séquence comme *équipements de modulation et de démodulation* est considérée comme équivalente à la séquence : *équipements de modulation et équipements de démodulation.*

Ces choix nous conduisent à extraire des termes de longueur > 2.

Deux automates ont donc été écrits : le premier regroupant les types élémentaires : N1 de (Det) N2 (*signal de raccrochage*), N1 à (Det) N2 (*tube à ondes*), N1 Prep N2 (*multiplexage par répartition*), N1 N2 (*voie support*)[3] ; le second pour le type élémentaire N Adj (*dissipation thermique*).

Une séquence morphosyntaxique reconnue par l'un des automates constitue une occurrence d'un couple appartenant à un de nos deux patrons : N Adj et N1 (Prep (Det)) N2. Un couple est constitué des deux lemmes qui correspondent aux deux extrémités lexicales de la séquence ; par exemple, pour le type élémentaire N1 (Prep (Det)) N2, le couple (satellite, orbite) pourra correspondre au séquences suivantes : *satellite sur orbite, satellites sur orbite, satellites en orbite, satellite mis en orbite*. Chaque séquence relevée est accompagnée de son schéma morphosyntaxique et de sa position dans le corpus (fichier, phrase, position dans la phrase).

3. Statistiques lexicales

ACABIT dans un deuxième temps utilise les résultats d'une évaluation de différentes mesures de statistique lexicale. Cette évaluation a permis de découvrir la meilleure mesure pour notre application, c'est-à-dire celle qui assigne un score élevée aux séquences les plus susceptibles de constituer des termes parmi notre liste de candidats.

3 Les parenthèses à l'intérieur des structures morphosyntaxiques indiquent le caractère optionnel d'une ou de plusieurs étiquettes grammaticales

Les caractéristiques numériques calculées par ACABIT ont chacune un rôle particulier : les fréquences sont les paramètres du critère d'association retenu ; le critère d'association mesure la force du lien entre les deux lemmes du couple. D'un point de vue statistique, les deux lemmes qui forment un couple sont considérés comme deux variables qualitatives dont il s'agit de tester la liaison. Les données se représentent sous la forme d'un tableau croisé, appelé tableau de contingence et défini à partir des comptes précédents.

Un tableau de contingence est associé à chaque couple de lemmes (L_i, L_j) :

	L_j	$L_{j'}$ avec $j' \neq j$
L_i	a	b
$L_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d résument les occurrences d'un couple :

a = le nombre d'occurrences du couple (L_i, L_j) ;
b = le nombre d'occurrences des couples où L_i est le premier élément d'un couple et L_j n'est pas le second ;
c = le nombre d'occurrences des couples où L_j est le second élément du couple et L_i n'est pas le premier ;
d = le nombre d'occurrences de couples où ni L_i ni L_j n'apparaissent.

La somme (a + b + c + d), notée N, est le nombre total d'occurrences de tous les couples trouvés pour un patron morphosyntaxique.

La littérature statistique regorge de mesures destinées à « tester l'indépendance » ou à « mesurer la liaison » ou encore à « mesurer le degré de similitude ou d'affinité » entre deux variables régies par un tableau de contingence. Dans Daille *et al.* (1995), nous avons évalué une dizaine de mesures dont : le score d'association, proche du concept d'information mutuelle, introduit par Church et Hanks (1990) :

$$IM = \log_2(a / (a+b)(a+c)) \qquad (1)$$

le coefficient du Φ^2 proposé par (Gale et Church, 1991) :

$$\Phi^2 = (ad - bc)^2 / (a+b)\,(a+c)\,(b+c)\,(b+d) \qquad (2)$$

ou encore le coefficient de vraisemblance présenté par (Dunning, 1993) :

$$\begin{aligned} Loglike = {} & a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) \\ & - (a+c) \log (a+c) - (b+d) \log (b+d) - (c+d) \log (c+d) \\ & + (a+b+c+d) \log(a+b+c+d) \qquad (3) \end{aligned}$$

Cette évaluation a démontré que la fréquence d'un couple est un très bon indicateur de son caractère terminologique. Ce résultat a contredit de nombreux travaux ré-

cents dans le domaine de l'extraction de ressources lexicales, qui proclamaient que l'information mutuelle donnait de meilleurs résultats que la fréquence (voir par exemple, Church et Hanks (1990)). Néanmoins, le classement proposé par la fréquence intégrant très rapidement du bruit, *i.e.* des couples qui ne réfèrent pas à des termes, nous avons choisi de ne retenir que le coefficient de vraisemblance (formule 3). Le coefficient de vraisemblance sélectionne les termes du domaine en leur attribuant une valeur forte : l'amplitude des valeurs dépend du nombre d'occurrences du couple ; plus le couple est fréquent plus la valeur du coefficient de vraisemblance tend à être élevée et ce indépendamment du nombre de couples extraits.

4. Expérimentation

ACABIT a été appliqué à deux corpus : *Le manuel des télécommunications par satellite (MTS)* (200 000 mots) et *Le livre bleu des communications (LBC)* (800 000 mots). Il en a extrait des couples pour les deux patrons N1 (Prep (Det)) N2 et N Adj. Une occurrence d'un couple correspond à une cooccurrence où les deux éléments du couple entrent dans un de ces deux patrons syntaxiques. Les tableaux 2 résument les fréquences des cooccurrences exprimées en nombre de couples ; ainsi pour le corpus *MTS* et le patron N Adj, nous avons relevé 4 483 couples dont 3 144 n'ont qu'une occurrence, 655 deux occurrences et 684 plus de deux occurrences.

MTS	1 occurrence	2 occurrences	plus de 2 occurrences	total
N Adj	3 144	655	684	4 483
N1 (Prep (Det))N2	6 834	1 503	1 616	9 953

LBC	1 occurrence	2 occurrences	plus de 2 occurrences	total
N Adj	5 201	1 507	2 113	8 821
N1 (Prep (Det))N2	12 167	3 481	6 288	21 936

TABLEAUX 2 : Nombres de cooccurrences extraites.

ACABIT applique ensuite aux couples, dont le nombre d'occurrences est au moins égal à deux, le critère d'évaluation retenu par notre évaluation. Nous obtenons donc en sortie de notre programme une liste ordonnée de couples ; chaque couple représentant un concept possible du domaine. Nous donnons dans les tableaux 3 et 4 qui concernent respectivement les patrons N1 (Prep (Det)) N2 et N Adj, les valeurs les plus élevées du coefficient de vraisemblance pour nos deux corpus. Nous utilisons les notations suivantes : Logl pour le coefficient de vraisemblance (*Loglike*) et Nbc pour le nombre d'occurrences du couple.

Sous chaque couple, nous trouvons toutes les variantes présentées en section 2.1.3. rencontrées dans nos corpus ainsi que leur localisation.

Couples de structure N1 (Prep (Det)) N2 dans ***MTS***	Séquence la plus fréquente	**Logl**	Nbc
(largeur, bande)	*largeur de bande* (197)	**1328**	223
(température, bruit)	*température de bruit* (110)	**777**	126
(bande, base)	*bande de base* (142)	**745**	145
(amplificateur, puissance)	*amplificateur(s) de puissance* (137)	**728**	137
(temps, propagation)	*temps de propagation* (93)	**612**	94
(règlement, radiocommunication)	*règlement des radiocommunications* (60)	**521**	60
(produit, intermodulation)	*produit(s) d'intermodulation* (61)	**458**	61
(taux, erreur)	*taux d'erreur* (70)	**420**	70
(mise, œuvre)	*mise en œuvre* (47)	**355**	47
(télécommunication, satellite)	*télécommunication(s) par satellite* (88)	**353**	99
(bilan, liaison)	*bilan(s) de liaison* (37)	**344**	55
Couples de structure N1 (Prep (Det)) N2 dans ***LBC***	Séquence la plus fréquente	**Logl**	Nbc
(canal, sémaphore)	*canal / canaux sémaphore(s)* (1188)	**5738**	1188
(accusé, réception)	*accusé de réception* (558)	**3983**	592
(système, signalisation)	*système(s) de signalisation* (82)	**2417**	85
(complément, étude)	*complément d'étude* (242)	**1985**	245
(point, sémaphore)	*point(s) sémaphore(s)* (677)	**1822**	679
(intervalle, temps)	*intervalle(s) de temps* (249)	**1782**	251
(trame, sémaphore)	*trame(s) sémaphore(s)* (354)	**1444**	354
(signal, fin)	*signal / signaux de fin* (385)	**1407**	391
(sous-système, utilisateur)	*sous-système utilisateur* (195)	**1226**	195
(bout, bout)	*bout en bout* (136)	**1155**	137
(contrôle, continuité)	*contrôle(s) de continuité* (171)	**1116**	171

TABLEAU 3 · Classement des couples de structure N1 (Prep (Det)) N2 proposé par ACABIT.

Couple de structure N Adj dans ***MTS***	**Logl**	Nbc	Couple de structure N Adj dans ***LBC***	**Logl**	Nbc
(station, ternen)	**2934**	750	(équipement, terminal)	**1425**	275
(débit, binaire)	**716**	134	(considération, général)	**1385**	25
(accès, multiple)	**605**	105	(service, supplémentaire)	**1275**	340
(voie, téléphonique)	**512**	118	(télégraphie, harmonique)	**1250**	152
(liaison, montant)	**457**	88	(étude, ultérieur)	**1171**	169
(liaison, descendant)	**408**	77	(caractère, graphique)	**1112**	19
(secteur, spatial)	**341**	79	(entité, fonctionnel)	**999**	19
(service, fixe)	**326**	66	(centre, international)	**964**	325
(lobe, latéral)	**299**	40	(adresse, complet)	**874**	183
(faisceau, hertzien)	**244**	35	(effet, local)	**865**	169
(puissance, surfacique)	**205**	35	(station, mobile)	**855**	164

TABLEAU 4 : Classement des couples de structure N Adj proposé par ACABIT

Rares sont les couples n'admettant aucune variante. Nous donnons ci-dessous quelques exemples de couples accompagnés des occurrences extraites :

(demande, trafic) : *demande de trafic, demandes en trafic, demande réelle en trafic.*

(liaison, satellite) : *liaison par satellite, liaisons par satellite, liaisons (très rapides + numériques + téléphoniques nationales) par satellite, liaisons numériques par satellites, liaisons satellite, liaisons entre satellites.*

(signal, fin) : *signal de fin, signaux de fin, signal (local + national + valide + périodique) de fin, signal émis à des fins, signal numérique utilisé à des fins*[4].

(ligne, abonné) : *ligne d'abonné, lignes d'abonné, ligne de l'abonné, lignes de l'abonné, ligne d'abonnés, lignes des abonnés, ligne(s) (téléphonique(s) + numériques(s) + analogique(s)) d'abonné, ligne(s) (numérique(s) + analogique(s)) de l'abonné, lignes et services d'abonné.*

Ces quelques exemples montrent que les modifieurs pouvant s'insérer à l'intérieur d'un terme binaire sont en nombre réduit. L'enregistrement de ces modifieurs, comme des autres altérations que subit la structure de base, y compris les différentes flexions rencontrées, n'est pas une tâche insurmontable surtout si celle-ci est effectuée automatiquement. Ces informations lexicales sont présentes sous l'entrée de chaque couple et pourront donc être directement intégrées dans une banque terminologique.

4 Les occurrences *signal émis a des fins* et *signal numérique utilisé à des fins* illustrent un problème de notre approche quantitative : nous n'avons aucune assurance qu'un couple regroupe des cooccurrences désignant un seul et unique concept ou encore, comme ici, des cooccurrences qui soient toutes valides.

5. Conclusion

Nous avons présenté une nouvelle approche pour l'extraction automatique de terminologies monolingues qui allie informations linguistiques et mesures statistiques. Cette méthode nous a permis d'extraire automatiquement un certain nombre de candidats termes du domaine classés selon leur pertinence terminologique à partir de corpus préalablement étiquetés et lemmatisés. Tous ces candidats termes sont accompagnés de leurs variantes morphologiques et de certaines de leurs variantes morpho-syntaxiques, ainsi que des modifieurs qui altèrent leurs structures. ACABIT permet donc une amélioration sensible des performances dans l'aide à la construction de banques terminologiques à partir de corpus et permet d'établir l'efficacité de l'enrichissement des systèmes statistiques par la linguistique. Cette méthode a été étendue à l'extraction de termes bilingues à partir de corpus alignés phrase à phrase (voir nos travaux dans Daille *et al.* (1994)).

Conception et exploitation d'un logiciel d'extraction de termes : problèmes théoriques et méthodologiques[1]

Didier BOURIGAULT

Centre d'analyse et de mathématiques sociales, (Unité mixte EHESS-CNRS-Paris Sorbonne), et EDF-Direction des études et recherches, Clamart, France

1. Introduction

La conception, la réalisation et l'utilisation d'un système automatique d'extraction terminologique conduisent à aborder sous un angle nouveau les questions théoriques et méthodologiques de la terminologie. Dans cet article, nous exposons l'état de notre réflexion sur quelques-unes de ces questions. Cette réflexion est à la fois exigée et nourrie par nos recherches sur la conception et la réalisation du logiciel d'aide au dépouillement terminologique Lexter (Logiciel d'EXtraction de TERminologie), ainsi que par les expériences d'utilisation effective de Lexter dans divers projets de recherche et développement à la Direction des études et recherches d'Électricité de France. Nous ne décrivons pas le logiciel en tant que tel dans cet article. Nous renvoyons le lecteur intéressé par les techniques de Traitement Automatique des Langues Naturelles (TALN) implémentées dans le logiciel à des publications antérieures (Bourigault, 1993a, 1993b, 1994, 1995).

La discipline terminologique hérite d'une définition du terme, établie dans le cadre de la Théorie Générale de la Terminologie fondée par Eugen Wüster en 1931, qui confère au terme le statut de symbole d'une notion. Même si elle fait parfois l'objet de débats, cette définition recueille un certain consensus dans la communauté des chercheurs en terminologie. Dans la section 2 de cet article, après un rapide survol historique (section 2.1.), nous exposons en quoi notre expérience de conception et d'utilisation d'un logiciel d'extraction de termes nous amène à participer à une critique constructive de cette définition (section 2.2.).

1 L'auteur remercie Benoît Habert (ELI, École normale supérieure de Fontenay Saint-Cloud) pour ses conseils et commentaires

Dans la section 3, nous donnons une caractérisation de la notion de *candidat terme*, en décrivant les critères de validité syntaxique et d'autonomie discursive, à partir desquels sont établies les règles opératoires de dépistage implémentées dans notre système automatique d'extraction de termes. La section 4 est consacrée aux aspects méthodologiques concernant la conception et la réalisation d'un logiciel d'extraction de terminologie.

2. Définition du terme : problèmes théoriques

2.1. Survol historique

En France, É. Benveniste et L. Guilbert sont parmi les premiers linguistes à s'être intéressés aux termes techniques des vocabulaires spécialisés. Dans son article sur la composition nominale, Benveniste (1966) entreprend de donner un statut à une nouvelle forme de composition, à la base de toutes les nomenclatures techniques, et il propose le terme nouveau de « synapsie ». Il caractérise la synapsie par un ensemble de sept traits, qui sont tous de type morpho-syntaxique, à l'exception du dernier, qui en appelle à une caractérisation sémantique de la synapsie : « le caractère unique et constant du signifié ». Cette dernière caractéristique est mentionnée au même niveau que les autres. Mais il apparaît clairement à la lecture de l'article que Benveniste la considère comme primordiale : « C'est toujours et seulement la nature du désigné qui permet de décider si la désignation syntagmatique est ou n'est pas une synapsie ». Cependant, Benveniste ne mentionne pas une éventuelle spécificité sémantique de la synapsie par rapport au mot de la langue. Il n'indique pas si la fonction de désignation complète et unique de la synapsie en fait une entité linguistique d'un type différent. Dans un article antérieur, où il complétait les thèses de Saussure sur l'arbitraire du signe, Benveniste (1939) avait insisté sur le fait que le lien entre le signifiant (image acoustique) et le signifié (représentation mentale) n'était pas arbitraire mais, au contraire, nécessaire. À la lecture conjointe de ces deux articles, il n'est pas aisé de cerner la position de Benveniste sur les compatibilités ou rapports entre le caractère nécessaire du lien entre signifiant et signifié dans le signe linguistique, le caractère arbitraire de la relation entre le signe linguistique et l'objet extralinguistique et la désignation complète et unique de l'objet par la synapsie.

Dans un travail de grande envergure, Guilbert (1965) décrit comment s'est formé le vocabulaire spécifique de l'aviation, entre les années 1861 et 1890. Il identifie et étudie de façon approfondie quatre formes de néologisme dans la formation du vocabulaire de l'aviation : néologisme morphologique, néologisme sémantique, néologisme grammatical, néologisme syntagmatique. Cette dernière forme, qui est la plus productive, contribue à la création d'unités lexicales complexes. La notion d'unité lexicale complexe recouvre partiellement la notion de synapsie de Benveniste, mais Guilbert propose une caractérisation plus détaillée, et différente sur certains points. Guilbert mentionne trois critères qui distinguent selon lui l'unité lexicale complexe du groupement syntagmatique du discours : la stabilité du rapport syntagmatique au plan du discours, la stabilité du rapport de signification entre l'unité syntagmatique et un signifié unique, la fréquence d'emploi qui stabilise à la fois le lien syntagmatique et le rapport de signification. Le second critère est, pour Guilbert, essentiel. À l'instar de Benveniste, il considère que la caractérisation essentielle de l'unité lexicale complexe est « d'essence sémantique ». La caractérisation de l'unité lexicale complexe, par op-

position au syntagme du discours, est la « constance » (ou la « permanence ») du rapport de signification entre l'unité syntagmatique et un signifié unique. Dans la conclusion de sa thèse, Guilbert adopte une position plus claire que celle de Benveniste sur la spécificité sémantique de l'unité lexicale complexe. La caractérisation qu'il en a proposée le conduit à mettre en doute la conformité du signe ainsi conçu avec le signe linguistique saussurien : le signe linguistique constitué par l'unité lexicale complexe « tendrait à se confondre avec le pur symbole qui n'a de contenu sémantique, toujours et en toutes circonstances, que celui qui lui a été préalablement conféré ».

De nos jours, la doctrine terminologique moderne prend appui sur les travaux fondateurs de Wüster développés dans le cadre de sa Théorie Générale de la Terminologie. Dans le cadre de la Théorie Générale de la Terminologie de Vienne, parue en 1931, Wüster avait proposé un modèle du terme qui tentait de concilier les théories de Saussure sur le signe linguistique et le triangle sémiotique « classique », proposé par un certain nombre d'auteurs et dont les sommets représentent le symbole, la notion et l'objet. Un certain consensus règne actuellement dans la communauté des terminologues sur la définition du terme. La définition donnée par H. Felber (1987), de l'École de Vienne, dans son manuel de terminologie, fait autorité : « Un terme est un symbole conventionnel représentant une notion définie dans un certain domaine de savoir ». Prise littéralement, cette définition du terme consacre une rupture entre le terme et le mot de la langue.

2.2. « Revisiter » la doctrine terminologique

Le point qui nous semble d'abord critiquable dans la définition du terme imposée par la doctrine terminologique est que cette définition participe d'une conception de la terminologie qui donne la primauté aux concepts sur leurs expressions linguistiques, ces dernières n'étant considérées que comme de simples symboles censés représenter de façon univoque ces notions. Cette vision mécaniste du couplage entre le terme et la notion s'est imposée dans le cadre intellectuel de l'universalisme et de l'empirisme logique, que le monde scientifique a depuis largement remis en cause (Slodzian, 1994, 1995). Cette vision ignore la complexité des phénomènes langagiers qui sont à l'œuvre dans les textes spécialisés, comme dans tous les types de textes, et donc interdit une analyse sémantique productive de ce type de textes. Nous souhaitons conjuguer nos efforts à ceux des chercheurs en linguistique, en socio-linguistique, en épistémologie, ainsi qu'en terminologie, qui visent à « revisiter » la doctrine terminologique, pour proposer une approche renouvelée de la terminologie qui intègre la terminologie au sein de la linguistique (de spécialité). Nos arguments prennent leur source dans la réflexion théorique qu'exigent la conception et la réalisation d'un outil automatique d'analyse de textes pour l'extraction de terminologie, ainsi que son utilisation dans des contextes applicatifs réels.

Sur le plan pragmatique, la conception doctrinaire du terme et de la terminologie se heurte à certaines réalités de la pratique terminographique, qui doit répondre à des besoins nouveaux. Ces besoins naissent de la demande plus forte pour une maîtrise des terminologies spécialisées et pour leurs utilisations dans des contextes de plus en plus diversifiés (Condamines, 1995). La normalisation n'est pas l'objectif prioritaire de l'activité terminologique dans les entreprises et les grandes organisations. Les besoins se situent d'abord au niveau de la description. Les types de « produits ter-

minologiques » que les terminographes vont avoir à réaliser sont de plus en plus variés. À côté des applications traditionnelles que constituent les bases de données terminologiques multilingues pour la traduction et les recueils de définitions pour l'enseignement, on voit apparaître de nouveaux types d'applications. C'est ainsi que, à la Direction des études et recherches d'EDF, nous menons diverses expériences dans lesquelles le logiciel d'aide au dépouillement terminologique Lexter est utilisé pour la réalisation de produits terminologiques de types divers. Ces projets s'inscrivent principalement dans le champ de la Gestion Électronique de Documents (GED). Nous concevons des systèmes de consultation de documentation technique, dans lesquels la terminologie joue un rôle central, sous la forme soit d'index terminologique structuré, soit de modèle terminologico-conceptuel.

Les besoins accrus en accès aux bases textuelles de plus en plus volumineuses feront certainement émerger la nécessité d'autres types de produits terminologiques. Pour toutes ces applications, le critère de définition du terme comme symbole d'une notion dans un domaine, déjà critiquable sur le plan théorique, s'avère non opératoire sur le plan empirique. En effet, l'expérience montre que, selon le type de produit terminologique à construire, les éléments lexicaux retenus, pour un même domaine et à partir d'un même corpus, seront différents.

Il convient donc d'abandonner l'approche néopositiviste qui pose la préexistence *a priori* d'un système de concepts que la terminologie aurait pour charge de dévoiler, pour adopter une démarche constructiviste et fonctionnelle plus propre à une approche linguistique de l'analyse des textes spécialisés, comme celle proposée par Lerat (1995). Il faut distinguer la notion théorique de *concept* dans les sciences cognitives, comme constituant fondamental de la pensée et des croyances, et la notion opératoire de *concept* en linguistique de spécialité. Dans le cadre qui nous concerne ici, à savoir la construction d'un modèle terminologico-conceptuel d'un domaine ou d'une activité, le concept est un construit. Il est le résultat produit par une analyse sémantique réglée d'un corpus constitué et pour une utilisation identifiée. Notre position sur ce point s'appuie sur les propositions de Rastier (1987, 1994, 1995). Construire un modèle terminologico-conceptuel, c'est produire un artefact, dans un code sémiotique particulier, celui des modèles de représentation des connaissances développés par l'intelligence artificielle. Le passage au concept est donc le résultat d'un travail de modélisation. La linguistique de spécialité doit s'intéresser aux conditions de ce passage, et ici elle doit se rapprocher de cette branche de l'intelligence artificielle qui s'intéresse à l'acquisition et à la modélisation des connaissances (Bourigault et Lépine, 1995 ; Bachimont, 1995).

3. La notion de candidat terme

3.1. De la notion de terme à celle de candidat terme

Le contexte de la conception et de la réalisation d'un logiciel d'aide au dépouillement terminologique est le suivant : il s'agit de concevoir un programme informatique qui recevra en entrée un corpus de textes techniques portant sur un domaine (quelconque), et qui devra en sortie proposer des mots ou des séquences de mots, extraites de ce corpus, qui pourront être retenus par un analyste humain en charge de la construction d'un produit terminologique (lui aussi quelconque). Nous imposons donc au système

une double contrainte de généralité, sur le domaine et sur le produit terminologique à construire.

La mise au point contrôlée d'un outil informatique d'aide au dépouillement terminologique exige une caractérisation linguistique rigoureuse du type de séquences, que nous désignerons sous le nom de « candidats termes », que ce système va avoir à extraire par analyse du corpus traité. La critique de la définition du terme, esquissée dans la section précédente, ne remet pas en cause la possibilité d'une telle caractérisation, mais au contraire la rend possible. Il s'agit de s'appuyer sur le constat selon lequel tout lecteur, averti ou non, est capable de relever dans les textes spécialisés des séquences de mots que son intuition pousserait à qualifier de « termes », ou, de façon plus prudente d'« unité lexicale complexe », d'« unité polylexicale », de « lexie complexe », ou encore de « synapsie ». Dans le cadre de la conception d'un outil d'aide au dépouillement terminologique, nous avons cherché à formaliser autant que possible les bases sur lesquelles repose ce jugement interprétatif, de façon à en déduire des règles opératoires de dépistage.

Nous caractériserons un candidat terme comme toute séquence *attestée dans le corpus* traité qui vérifie les contraintes de *validité syntaxique* et d'*autonomie discursive*. Cette caractérisation est sujette à discussion. Comme nous le verrons dans les deux sections suivantes, aucun des deux critères mentionnés ne va sans poser de sérieuses questions, à la fois sur le plan théorique de l'analyse linguistique et le plan pratique de l'implémentation informatique. Néanmoins, l'utilité de cette caractérisation est de fournir un cadre d'analyse pour guider la recherche de règles opératoires, implémentables dans un système d'analyse automatique de textes.

3.2. Le critère de la validité syntaxique

Le premier critère de caractérisation du candidat terme est celui de la validité syntaxique : un candidat terme doit correspondre à une séquence syntaxiquement valide dans la proposition de laquelle il a été extrait. Nous illustrons ce critère à l'aide des quelques exemples donnés dans le tableau 1.

L'exemple 1 pose le problème classique du rattachement des adjectifs et des groupes prépositionnels. La description syntaxique du groupe révèle que la séquence **file de filtration** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, bien qu'elle corresponde à un patron terminologique valide ('nom préposition nom'). Le sous-groupe **filtration iodée** est un bon candidat terme.

L'exemple 2 illustre le problème posé par les propriétés de sous-catégorisation des adjectifs. La description syntaxique du groupe révèle que la séquence **capteur sensible** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, bien qu'elle corresponde à un patron terminologique valide ('nom adjectif'). Cela ne signifie pas que la collocation qui associe l'unité lexicale « capteur » à l'unité lexicale « sensible » ne soit pas digne d'intérêt, mais pour conserver un contrôle sur les traitements effectués par le système, il est nécessaire de contraindre celui-ci à ne chercher à extraire que des séquences syntaxiquement valides.

description syntaxique :	[1]	*file de filtration iodée* [file de [filtration iodée]] (-) **file de filtration**
description syntaxique :	[2]	*un capteur sensible à une élévation de température.* un [capteur [sensible à une élévation de température]] (-) **capteur sensible**
description syntaxique :	[3]	*On installe le câble contre le coffret de décharge.* On installe [le câble] [contre le coffret de décharge] (-) **câble contre le coffret de décharge**
description syntaxique :	[3bis]	*On installe une protection contre les grands froids.* On installe [une protection contre les grands froids] (+) **protection contre les grands froids**

TABLEAU 1 : Illustration du critère de la validité syntaxique. Sous chaque exemple figure sa description syntaxique (partielle) correcte On en déduit les séquences qui vérifient (+) ou ne vérifient pas (-) le critère de la validité syntaxique

Les exemples 3 et (3bis) illustrent le problème posé par les propriétés de sous-catégorisation des noms. La description syntaxique de l'exemple (3) révèle que la séquence **câble contre le coffret de décharge** n'est pas un bon candidat terme au sens du critère de la validité syntaxique, alors que dans l'exemple (3bis), qui présente une configuration analogue en terme de succession de catégorie grammaticale (un nom suivi de la préposition « contre » et de l'article défini), la description syntaxique révèle que la séquence **protection contre les grands froids** constitue elle un bon candidat terme. Dans ce dernier cas, le nom « protection » est construit avec un complément introduit par la préposition « contre », alors que dans l'exemple (3), cette préposition introduit un complément du verbe « raccorde » de la proposition principale.

3.3. Le critère de l'autonomie discursive

Le second critère de caractérisation du candidat terme est celui de l'autonomie discursive : un candidat terme doit posséder une autonomie discursive vis-à-vis du contexte textuel duquel il a été extrait. Ce critère vient réduire encore le champ des possibles. Parmi les groupes syntaxiquement valides, il faut éliminer ceux qui n'ont pas d'autonomie contextuelle, c'est-à-dire ceux qui ne peuvent être interprétés sans le contexte textuel antérieur ou postérieur. Ceux-là ne peuvent en aucun cas être extraits de leur contexte textuel pour être intégrés comme entrée autonome dans un quelconque produit terminologique.

Nous illustrons ce critère à l'aide des quelques exemples donnés dans le tableau 2, qui concernent tous l'article défini en français.

En discours, l'article défini est passible de diverses valeurs sémantiques, qui ne

sont pas toutes compatibles avec une extraction hors contexte. Dans les exemples (4) et (5), l'article défini, avec les valeurs anaphorique et cataphorique ne peut être correctement interprété sans l'accès aux contextes textuels antérieur et postérieur. Les séquences **intégrité de la paroi** ne sont pas de bons candidats termes au sens du critère de l'autonomie discursive.

Par contre, dans l'exemple (6), l'article défini a la valeur d'unique. Le corpus traité constitue la description d'une tranche nucléaire générique, d'un type bien défini. Dans ce cas, le groupe « la tranche » supporte toujours la même interprétation (« la tranche générique » dont on donne la description dans cette documentation), et peut donc être extrait de tout contexte textuel en tant que constituant de candidat terme. Dans cet exemple, la séquence **niveau de puissance de la tranche** est un bon candidat terme au sens du critère de l'autonomie discursive. L'exemple (7) illustre la valeur de générique qui elle aussi est compatible avec une extraction hors contexte.

[4] *Le circuit d'aspersion est installé contre une paroi de confinement. Il a pour rôle de maintenir l'intégrité de* ***la*** *paroi.*	
valeur de l'article défini :	anaphorique
	(-) **intégrité de la paroi**
[5] *Le circuit d'aspersion a pour rôle de maintenir l'intégrité de* ***la*** *paroi contre laquelle il est installé.*	
valeur de l'article défini :	cataphorique
	(-) **intégrité de la paroi**
[6] *Ce système règle le niveau de puissance de* ***la*** *tranche.*	
valeur de l'article défini :	unique
	(+) **niveau de puissance de la tranche**
[7] *sensibilité à* ***la*** *chaleur*	
valeur de l'article défini :	générique
	(+) **sensibilité à la chaleur**

TABLEAU 2 . Illustration du critère de l'autonomie discursive. Sous chaque exemple figure la valeur sémantique de l'article défini. On en déduit les séquences qui vérifient (+) ou ne vérifient pas (-) le critère de l'autonomie discursive.

4. Aspects méthodologiques de conception : approche expérimentale

Étant donné la caractérisation que nous venons de donner du candidat terme, il apparaît clairement qu'un système automatique de repérage de candidats termes doit mettre en œuvre une analyse de type syntaxique, comme l'ont déjà souligné David et Plante (1990). Le critère de la validité syntaxique exige que le système soit en particulier doté de règles d'analyse capables de résoudre au mieux les problèmes de rattachement dans les situations ambiguës. Le critère de l'autonomie pose des problèmes

d'implémentation encore plus délicats. L'essentiel de nos efforts actuels de développement du logiciel Lexter concerne ce point. Nous ne décrivons pas dans cette section les techniques de Traitement Automatique des Langues Naturelles mises en œuvre dans Lexter, nous exposons quelques points de méthode concernant les phases de conception et de réalisation d'un tel système.

Il s'agit d'implémenter des règles de dépistage qui extraient du corpus d'apprentissage des séquences de mots qui satisfassent autant que possible les critères de la validité syntaxique et de l'autonomie discursive. La tâche de conception et de réalisation d'un système automatique de dépistage de candidats termes exige que soient menées de front deux types d'activité : une activité de recherche théorique sur la caractérisation linguistique du terme et sur son fonctionnement discursif, et une activité d'implémentation informatique de règles de dépistage dans l'outil d'analyse automatique. Ces deux activités se développent conjointement dans une démarche dialectique très fructueuse. L'analyse théorique guide la réalisation informatique, tout en profitant de ses résultats. Les deux activités se nourrissent mutuellement. La linguistique joue un rôle de régulation, avant, pendant et après la réalisation informatique :

- en amont de la conception du système, une analyse linguistique du fonctionnement du terme en discours est nécessaire pour établir les principes de base optimaux de l'analyse automatique ;
- pendant la mise au point du système, les règles d'analyse effectivement implémentées sont établies dans une démarche linguistique expérimentale privilégiant le test sur corpus ;
- en aval, la validation de l'outil et l'édification d'une méthode d'utilisation de l'outil par un expert humain procède encore d'une analyse linguistique et ergonomique de l'activité terminologique.

L'activité de dépouillement terminologique est une activité d'analyse conceptuelle d'un domaine, et donc en ce sens une activité « hautement » intellectuelle. Doter une machine de règles de dépistage de candidats termes dans un texte relève d'une certaine gageure. Mais l'analyse linguistique, à laquelle se subordonne l'implémentation informatique, permet de gérer et de maîtriser le processus d'approximation que constitue l'établissement de règles opératoires pour une machine. En ce sens, elle révèle et assume les limites des capacités d'une machine à travailler sur le sens.

5. Conclusion : linguistique de corpus

La démarche de conception d'un logiciel d'aide au dépouillement terminologique est donc par essence de type expérimental. Analyse théorique, investigation et expérimentation sur corpus sont menées de pair. Le corpus joue un rôle de pivot dans la démarche :

(i) en tant qu'objet d'analyse pour le système ;
(ii) en tant que source d'information pour le système (Lexter est doté de procédures dites « d'apprentissage endogène » qui lui permettent d'acquérir par lui-même certaines informations syntaxiques de sous-catégorisation dont il a besoin pour effectuer une analyse syntaxique précise) ;
(iii) en tant qu'élément de base du dispositif expérimental.

Ce dernier aspect est essentiel. Une fois donnés les principes généraux de conception, nous avons progressivement élaboré les techniques d'analyse et les règles des différents modules du système en associant réflexion linguistique et validation par test sur corpus. L'analyseur sert d'outil d'investigation dans les corpus (on peut parler dans ce cas d'analyse linguistique assistée par ordinateur). Il s'agit alors de concilier les visées de l'analyse linguistique qui met en exergue les phénomènes marginaux et donne au contre-exemple un pouvoir de remise en cause, et les contraintes de la réalisation informatique, qui privilégient les phénomènes de masse. L'expérimentation sur corpus est une activité qui exige patience et rigueur, et qui peut être parfois fastidieuse. Parce qu'elle dévoile toujours des problèmes nouveaux, la confrontation avec le corpus est à la fois décourageante et passionnante.

Amélioration automatique incrémentale de dictionnaires bilingues utilisant un corpus monolingue

Kumiko TANAKA et Violaine PRINCE

Université de Tokyo, Japon et LIMSI-CNRS, Paris, France

1. Introduction

Pour développer automatiquement des dictionnaires électroniques bilingues, il faut aligner des corpus bilingues. Mais, réciproquement, l'alignement de corpus ne peut se faire sans l'aide de dictionnaires bilingues relativement complets (Utsuro *et al.*, 1994). Cette interdépendance entre l'alignement des corpus et la construction de dictionnaires électroniques bilingues est une des raisons pour lesquelles l'identification et la mise à jour de dictionnaires bilingues reste un problème difficile. L'objectif que nous nous proposons d'atteindre par le biais de l'algorithme présenté dans cette contribution est de transformer le problème de la mise en correspondance bilingue, dont la difficulté vient d'être citée, en un problème monolingue de recherche d'un ensemble de mots (M) sémantiquement proches de l'entrée lexicale originelle. Ce résultat peut ensuite être transféré dans le cadre bilingue, par transfert des équivalents des mots de M comme équivalents de l'entrée originelle.

Dans des travaux préalables de génération de dictionnaires bilingues par l'intermédiaire d'une troisième langue jouant le rôle de pivot, Tanaka et Umemura (1994) considèrent que les mots ayant des sens multiples dans la langue pivot transportent des équivalents qui ne sont pas des candidats pertinents. Néanmoins, ces candidats parasites peuvent être écartés en procédant à une consultation inverse du dictionnaire. Cette hypothèse peut ici être exprimée différemment : le problème de la correspondance de mots entre langues peut être rapporté à un problème intralinguistique (dans la langue pivot), problème de mesure de proximité sémantique entre deux mots. Dans ce même ordre d'idée, nous nous proposons – au lieu d'utiliser une troisième langue pivot et de transporter des termes parasites pour les écarter ensuite – de nous appuyer sur une étude de proximité sémantique grâce au traitement d'un corpus monolingue d'une part, et aux informations de synonymie et de proximité morphologique conte-

nues dans des lexiques électroniques de la langue source (LS) d'autre part, dans le but de compléter automatiquement des dictionnaires bilingues.

Pour cela, nous nous donnons trois types d'informations monolingues :

- des heuristiques calculant une certaine proximité sémantique grâce à la présence de similarités morphémiques (très indicatives de l'identité de racine en japonais) ;
- la présence de synonymes fournis par le lexique de la LS ;
- des valeurs de cooccurrence établies à partir de grands corpus (aussi en LS).

En pratique, le fait d'utiliser les corpus comme sources de relations sémantiques donne un aspect incrémental à notre algorithme de raffinement de dictionnaires. En effet, les corpus rendent compte des évolutions sémantiques et de la dynamique de la langue, ce qui est important pour la mise à jour des dictionnaires électroniques bilingues. Nous pensons que lorsque les dictionnaires sont incomplets ou comportent des équivalences qui ne sont plus forcément à jour, des extraits actualisés de discours que sont les corpus contemporains sont probablement les sources les plus riches pour transformer ces dictionnaires. De plus, il existe de nombreuses méthodes relativement éprouvées de calcul de la similarité sémantique dans les corpus, ce qui en fait une source aisément exploitable. Enfin, si les corpus sont spécialisés dans un domaine, on peut se servir de cette particularité pour spécifier plus précisément la terminologie en vigueur afin de produire des dictionnaires de spécialité.

Si nous avons choisi le japonais comme LS et l'anglais comme langue cible (LC), c'est pour les raisons suivantes :

- le japonais utilise les idéogrammes kanji et la formation de mots à partir d'idéogrammes se fait de telle sorte que la reconnaissance des racines morphologiques est évidente à mettre en œuvre et fournit beaucoup de renseignements ;
- le japonais est très différent des langues indo-européennes et on ne peut pas jouer sur la proximité des lexies pour deviner le sens, comme on peut le faire entre les langues à base latine ;
- un bon dictionnaire électronique bilingue japonais-anglais existe, ce qui permet de tester l'algorithme dont le but est d'être ensuite appliqué au français pour lequel il n'existe malheureusement pas de bons dictionnaires bilingues avec le japonais. Le bon dictionnaire servira de structure témoin, et pour simuler une situation de dictionnaire incomplet, nous « dégraderons » le dictionnaire d'origine (c'est-à-dire que nous en produirons une version délibérément amoindrie) de manière à voir dans quelle mesure l'algorithme et l'usage de corpus nous permettent d'au moins restaurer le bon dictionnaire d'origine. Par la suite, si l'algorithme s'avère bon, il sera facile de l'appliquer au dictionnaire japonais-français, qui sera effectivement enrichi par cette manœuvre.

Dans la section suivante, nous expliquons la méthode sous-jacente à notre algorithme en l'illustrant par un exemple. La section 3 reprend l'algorithme de manière formelle et montre le principe incrémental de la méthode. La section 4 décrit les données ; et la section 5 analyse les résultats obtenus.

Dans ce qui vient, les lexies japonaises sont translittérées en romain (alphabet romain) en *italique* avec chaque idéogramme kanji séparé par un tiret («-»). Nous avons

associé le sens de chaque forme translittérée entre parenthèses. Les termes anglais sont en courier. Les traductions françaises des termes anglais sont fournies entre parenthèses consécutivement à ces derniers.

2. Présentation générale de l'algorithme

Pour illustrer notre méthode nous avons pris le cas du mot japonais *ken-kyuu* (recherche). Nous avons commencé par produire un dictionnaire dégradé japonais-anglais, c'est-à-dire une dictionnaire dans lequel la lexie *ken-kyuu* n'est rattachée qu'aux deux mots anglais research (recherche) et work (travail). Nous souhaiterions pouvoir récupérer des entrées telles que search (recherche dans le sens de quête, ou enquête), investigation qui sont proches du terme anglais research (termes qui sont en fait des synonymes du mot français « recherche »).

Une manière de faire consiste à utiliser un corpus aligné et de compter les mots qui cooccurrent, dans les deux langues, avec *ken-kyuu*. Une autre manière consiste à s'appuyer sur des renseignements en provenance d'un thésaurus dans la LS. À partir d'un dictionnaire japonais, nous avons obtenu les informations suivantes : *ken-kyuu* est relié aux mots *chou-sa* (investigation) et *tan-kyuu* (quête, enquête). Nous pouvons donc raisonnablement penser que, dans un dictionnaire de qualité moyenne, *chou-sa* a des équivalents tels que search et investigation et que *tan-kyuu* serait relié avec research et search. Ces associations sont représentées dans le graphe de la figure 1.

À partir de cela, nous définissons la notion de **correspondance** entre deux mots, comme étant tout arc du graphe permettant de relier un mot avec un autre. Les équivalents d'une entrée lexicale peuvent être redéfinis comme des mots dans la LC, reliés à elle par des correspondances. Nous définissons de même la notion de **similarité**, comme étant la relation entre les mots de la LS en correspondance avec notre entrée lexicale. Ces mots sont alors appelés des **similaires**.

Dans notre exemple, le mot anglais search possède deux mots en correspondance avec lui, chacun passant par le biais de *chou-sa* d'une part, et de *tan-kyuu* d'autre part. De plus, le mot investigation possède une correspondance passant par le biais de *chou-sa*. De ce fait, le mot anglais search devrait avoir un lien plus fort avec *ken-kyuu* que n'en a le mot investigation[1]. Il faut aussi remarquer que dans la mesure où research reçoit des arcs aussi bien depuis *tan-kyuu* que depuis *ken-kyuu*, dès lors, *ken-kyuu* est plus fortement lié à research que dans la version originelle.

La manière la plus simple de définir l'importance d'une relation entre une entrée lexicale et ses équivalents consiste à pondérer les mots qui sont en correspondance avec elle. Par conséquent, nous appellerons le poids d'une correspondance (PC) entre deux mots *a* et *b* une valeur qui sera désignée par la formule *w(a, b)*.

1 Ce qui voudrait dire qu'il faudrait privilégier l'équivalent français « recherche » à l'équivalent français « investigation »

Les PC entre une entrée lexicale et ses équivalents sont initialement calculés à partir du dictionnaire dégradé. Ce poids est soit égal à l'inverse du nombre d'équivalents, ou bien peut dépendre de l'ordre des équivalents apparaissant dans le dictionnaire (on fait l'hypothèse que l'équivalent le plus fort est celui qui apparaît en premier, et ainsi de suite). Les valeurs données dans la figure 1 illustrent notre mode de calcul de *w(a, b)*.

Les PC entre l'entrée lexicale et ses similaires peuvent être judicieusement calculés à l'aide de trois heuristiques :

- la similarité morphologique (graphèmes communs en japonais) ;
- la coprésence de synonymes dans le thésaurus en LS ;
- les valeurs de cooccurrence de termes, valeurs obtenues dans de grands corpus en LS.

En ce qui concerne les aspects morphologiques, *tan-kyuu* possède un graphème (qui est aussi un morphème) commun avec *ken-kyuu* qui est *kyuu*. Nous définissons alors ce que nous appelons le **score morphologique** comme étant le nombre de kanjis (morphographèmes) communs entre deux termes de la LS. Par exemple, lorsque l'on compare *chou-sa* et *gaku-jutsu-chou-sa* il existe deux kanjis communs *chou* et *sa*, et donc le score morphologique est de 2. Dans le cas de notre exemple, la première ligne de la table 1 est ainsi obtenue.

Le deuxième nombre signifiant est le nombre de synonymes de l'entrée lexicale considérée, dans le lexique en LS. Dans notre expérience, *ken-kyuu* possède deux synonymes : *tan-kyuu*, *gaku-mon*. Dès lors, le poids de *tan-kyuu* augmente, mais pas celui de *chou-sa*. C'est ainsi que nous avons calculé les valeurs indiquées dans la deuxième ligne de la table 1.

Le troisième nombre que nous considérons est une valeur de cooccurrence dans un corpus. La manière la plus simple d'évaluer la cooccurrence est d'utiliser le principe d'information mutuelle de Church[2], valeur obtenue à partir de corpus en LS. La quatrième ligne de la table 1 est calculée à partir d'un corpus de 33 méga-octets, avec une fenêtre de 1 024 octets.

Les deux premiers nombres dénotent des informations statiques, alors que le troisième varie en fonction de la taille et de la nature du corpus.

En terme de méthode, nous commençons par transformer le dictionnaire à l'aide des poids d'origine statique, puis nous le transformons à nouveau à partir du poids en provenance du corpus. Ensuite nous normalisons chaque ligne de la matrice des poids en divisant chacune des valeurs qu'elle contient avec la somme des valeurs (norme

2 La valeur d'information mutuelle (Church, 1990) entre deux mots *a* et *b* (tous deux en LS) est obtenue par la formule suivante

$$\log_2 \frac{P(a, b)}{P(a)P(b)}$$

où P(a) est la probabilité de trouver le mot *a* dans le corpus

pondérée). C'est ainsi que nous engendrons les valeurs des lignes 1 et 3 de la matrice 1. Les lignes ainsi obtenues sont ensuite additionnées, et les sommes sont stockées dans la quatrième ligne, laquelle est de nouveau normalisée, et la valeur normale est stockée dans la cinquième ligne. Cette dernière ligne reflète les PC des similaires d'une entrée lexicale dont une illustration est fournie en figure 1.

Le raffinement du dictionnaire peut être défini comme la réévaluation des PC pour chaque équivalent d'une entrée lexicale. Cela peut se faire à partir de la somme des PC originels, *modulo* les poids obtenus par les trois scores (morphologique, synonymes et corpus), et cela, par le biais de la procédure suivante :

$$w_{new}(x_i, y_j) = r(w(x_i, y_j) + \sum_{\forall k \in D} w(x_i, x_k) w(x_k, y_j)) \quad (1)$$

où $r = 1.0/\sum w_{new}(x_i, y_j)$. Par exemple :

w(*ken* — *kyuu*, `research`)
= *r* x (0.9 + 0.75 x 0.2 + 0.25 x 0.6)
w(*ken* — *kyuu*, `work`) = *r* x 0.1
w(*ken* — *kyuu*, `search`) = *r* x (0.25 x 0.4 + 0.75 x 0.3)
w(*ken* — *kyuu*, `investigation`) = *r* x (0.75 x 0.8)

où *r* est la somme de *w*(*kenkyuu*, `research`), *w*(*kenkyuu*, `work`) et *w*(*kenkyuu*, `search`). Le nouveau graphe avec les nouveaux PC des équivalents de *ken-kyuu* est représenté dans la figure 2. Le dictionnaire s'est enrichi de nouvelles correspondances, dans la mesure où ce nouveau graphe possède plus d'arcs et de nœuds que le précédent, et cela, à partir de connaissances en LS.

Pour modifier les PC avec les poids obtenus à partir du corpus, nous appliquons la même procédure que pour les scores statiques, en utilisant les PC de la table 1 associés aux similaires de l'entrée lexicale. Le PC pour chaque paire de mots, dans une langue donnée, est impossible à calculer, puisque théoriquement la matrice des co-occurrences potentielles est de dimension 10^9. Par conséquent, pour restreindre la charge calculatoire, nous avons choisi de retenir les lexies japonaises qui appartiennent à au moins un des deux ensembles suivants :

- l'ensemble des synonymes de toutes les entrées, lexicales que nous considérons, synonymes se trouvant dans le lexique en LS que nous possédons ;
- les termes japonais obtenus en recherchant dans le dictionnaire anglais-japonais pour obtenir les équivalents anglais.

	tan-kyuu	*chou-sa*
morphème	1,00	0,00
synonyme	1,00 0,50	1,00 0,50
morph + syn	1,50 0,75	0,50 0,25
corpus	11,18 0,57	8,48 0,43

TABLEAU 1 · Nombres discriminants pour le poids de correspondance.

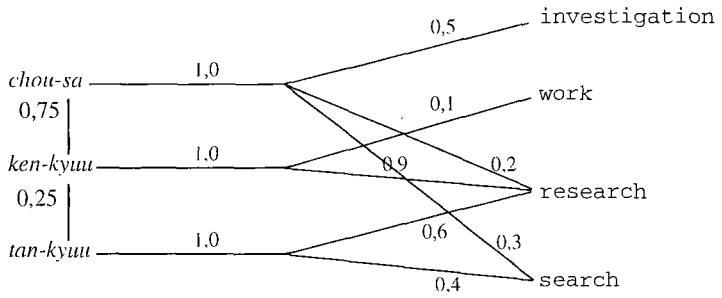

FIGURE 1 : Structure représentant les connaissances d'origine

En conclusion de cette présentation générale de notre méthode, nous dirons, pour résumer, que l'algorithme proposé réduit la problématique bilingue (c'est-à-dire « est-ce que deux mots dans des langues différentes ont un sens similaire ou pas ? ») en une problématique monolingue (c'est-à-dire « est-ce que deux mots, dans une même langue, ont un sens similaire ou pas ? »). Le problème ainsi transposé peut être ensuite restitué dans le cadre bilingue d'origine en recalculant les poids de correspondance entre une entrée lexicale en LS et ses équivalents en LC, ces poids étant modifiés par des scores dénotant des connaissances morphologiques et sémantiques (synonymes), puis normalisés. Dans la section suivante, nous décrivons formellement les différents pas de la méthode que nous avons suivie.

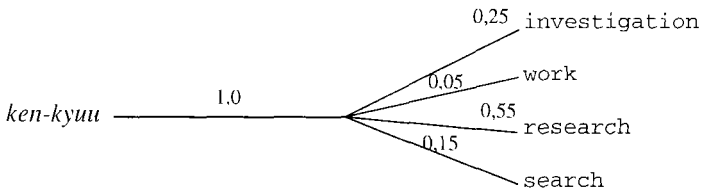

FIGURE 2 : Poids des correspondances affectés par les connaissances morphologiques et sémantiques.

3. Description formelle de la méthode

3.1. Matrices de correspondance bilingue et monolingue

Les mots de la LS sont notés *s* et ceux de la LC sont notés *t*. Nous définissons une matrice de correspondance bilingue, symbolisée par *BCM* (pour *Bilingual Correspondence Matrix*) dont chaque (*i*, *j*)-th élément est noté $w(s_i, t_j)$. Comme nous l'avons précédemment expliqué, la matrice est initialisée à partir du dictionnaire bilingue d'origine, dont on sait qu'il est de qualité médiocre (incomplet ou éventuellement erroné). Dans cette matrice, on fait l'hypothèse que le mot s_i possède *n* équivalents notés t_k (k=1,...,n). Il existe deux façons de calculer la pondération initiale des équivalents. Soit on décide d'affecter une équiprobabilité à l'ensemble des termes, ce qui donnera une valeur initiale de $1/n$ à chaque élément $w(s_i, t_j)$, soit on décide de tenir compte de l'ordre de présentation des équivalents dans le dictionnaire initial et on affecte au *i*ème équivalent un poids égal à $1-(2i/(n(n+1)))$ (la somme d'un rang est 1,0).

Nous définissons ensuite une matrice de cooccurrence monolingue, symbolisée par *MCM* (pour *Monolingual Cooccurrence Matrix*) dans laquelle chaque élément (*i*, *j*)th est noté $w(s_i, s_j)$. Cet élément désigne les valeurs de pondération des similaires obtenues à partir des trois nombres discriminants représentant des connaissances lexicales (morphologiques, sémantiques et cooccurrence) tels que nous les avons présentés dans la section précédente.

Cette même section avait décrit la procédure locale de transposition de la problématique bilingue en monolingue, et de restitution dans le cadre d'origine, en l'illustrant par l'exemple de l'entrée lexicale japonaise *ken-kyuu*. En pratique, cette procédure se définit comme une multiplication de matrices et ce, par le biais de la formule suivante :

$$(MCM + E) \times BCM$$

E est la matrice unité (neutre de la multiplication). Elle est ajoutée pour bien marquer l'état originel de la matrice de correspondance bilingue, c'est-à-dire les correspondances originelles entre une lexie source *s* et une lexie cible *t*. Cette multiplication des correspondances est exactement ce qui est montré dans la formule 1.

Notons que l'addition et la multiplication de matrices sont toujours suivies par une normalisation, puisque les valeurs correspondent à des poids. Remarquons de même que la matrice résultante a la même dimension que *BCM*. Certains éléments ini-

tialement évalués à 0 (simplement parce qu'ils n'étaient pas représentés dans la pondération d'origine) reçoivent ensuite des poids non nuls, par addition des PC des équivalents.

3.2. Incrémentalité

Parmi les heuristiques considérées, nous avions déjà mentionné que le score morphologique et le nombre des synonymes étaient des informations statiques et relativement stables, alors que les valeurs de cooccurrence de termes dans des corpus apparaissent comme dynamiques et sans stabilité *a priori*, puisqu'elles peuvent varier en fonction de la nature du corpus et de sa taille. La matrice *BCM* obtenue après pondération par le biais du score morphologique et du nombre des synonymes est notée MCM_s. La matrice obtenue après pondération par le score de cooccurrence est notée MCM_c, où le terme *c* représente le corpus considéré.

Dans la précédente section, nous avons montré que cette transformation matricielle était réalisée en deux étapes. Effectivement, la première étape consiste à calculer MCM_s, c'est-à-dire à incrémenter le dictionnaire bilingue (la matrice *BCM* d'origine) à partir des sources d'information monolingues et statiques. La deuxième étape calcule la matrice MCM_c à partir de *BCM* et des valeurs de cooccurrence dépendant des propriétés du corpus *c*. En d'autres termes, cette dernière matrice va ajouter à la matrice origine des relations sémantiques dérivées d'un usage actuel de la LS.

Le raffinement final du dictionnaire s'obtient alors par la formule :

$$(MCM_c + E) \times (MCM_s + E) \times BCM$$

En pratique, quand nous parlons de processus incrémental, c'est pour désigner explicitement notre méthode d'obtention de la matrice MCM_c. Celle-ci n'est pas calculée directement, mais de manière récursive sur les corpus eux-mêmes, et ce, dans le but d'atteindre tout de même un minimum de stabilité des relations sémantiques dans les corpus. Il existe deux manières de faire varier l'impact du corpus et de réaliser l'incrémentation :

- soit on considère que la matrice MCM_c utilisée dans la procédure de raffinement définie ci-dessus est elle-même la résultante (le produit) de plusieurs matrices MCM_{ci} et donc qu'elle se définit comme la matrice obtenue sur l'ensemble des corpus d'expérience. Dans ce cas, le membre $(MCM_c + E)$, où *c* représente l'union de tous les corpus, de la formule de raffinement n'est multiplié qu'une seule fois ;
- pour chaque corpus d'expérience, nous calculons un *MCM* local, et ce dernier est appliqué incrémentalement au *BCM* courant, dans la formule de raffinement. Ce qui fait que nous obtenons des dictionnaires plus ou moins raffinés selon les corpus. Le développement de la formule donne :

$$(MCM_{c1} + E) \times \ldots \times (MCM_{c0} + E) \times (MCM_s + E) \times BCM$$

Dans les deux cas, nous n'avons pas la possibilité de démontrer qu'il existe une

distributivité de la loi de composition sur les corpus, c'est-à-dire que :

$$MCM_{c} \times MCM_{c'} = ?MCM_{(c+c')}$$

Cette distributivité voudrait dire que les corpus ont un effet additif, ce qui n'est pas une hypothèse forcément raisonnable en terme de relations de cooccurrence tout au moins. Nous avons choisi, pour cette première expérience, de réaliser une incrémentation du premier type, c'est-à-dire avec une matrice réalisée sur l'ensemble des corpus et multipliée une seule fois, en se réservant pour un futur proche la mise en œuvre du second type et la comparaison des deux techniques. Il est certain que l'incrémentalité à partir de corpus peut produire des effets secondaires : si, par exemple, le dictionnaire d'origine est très pauvre (comme ce sera malheureusement le cas pour notre dictionnaire japonais-français pour lequel toute notre expérience a été menée), et s'il est entraîné incrémentalement sur des corpus homogènes et spécifiques, alors il aura tendance à refléter cette spécificité. Par exemple, si nous entraînons notre dictionnaire dégradé avec des corpus scientifiques uniquement, il est très probable que nous obtiendrons une version scientifique du dictionnaire, ce qui dans certains cas, peut devenir fort utile pour créer des dictionnaires bilingues de spécialité.

corpus	occurrences	couverture
corpus6	3005549	49173
corpus5	1200000	45023
corpus4	672378	39850
corpus3	300000	27476
corpus2	150000	19883
corpus1	75000	13841

TABLEAU 2 : Corpus utilisés dans l'expérience.

4. Les données de l'expérience

Les dictionnaires japonais-anglais et anglais-japonais qui nous ont servi ici sont Ichikawa (1990) et Koine (1990). Le nombre d'entrées lexicales dans le japonais-anglais est 47 808 et nous en comptons 28 168 dans la version anglais-japonais. Le lexique électronique japonais provient de Takebe *et al.* (1976). Il contient 27 145 entrées. Pour les besoins de l'expérience, nous avons créé le dictionnaire dégradé de la manière suivante : les noms communs sont extraits de façon semi-automatique depuis chacun des deux dictionnaires bilingues et sont inclus dans un dictionnaire qui va grosso modo inclure des correspondances un à un entre les termes. De plus, le dictionnaire comprend des erreurs (contresens, absurdités).

Les corpus que nous avons choisis sont extraits des archives du journal *ASAHI* (l'équivalent japonais du journal français *Le Monde*) et forment environ 33 mégaoctets de texte. Le corpus de notre expérience correspond à la totalité du corpus dis-

ponible. Les données sont lemmatisées par le système JUMAN, et nous recueillons les noms et les verbes. Le nombre d'occurrences de chaque lexème retenu et le taux de couverture linguistique sont fournis dans le tableau 2. Pour fournir un cadre à la fonction incrémentale citée dans la section précédente, nous avons coupé l'ensemble des données en six (numérotés de 1 à 6). Les corpus 1 à 5 sont obtenus à partir des différentes parties principales du corpus6 (qui, lui, correspond à la totalité du texte). Par exemple, corpus5 correspond à la première partie de corpus6 et corpus4 à celle de corpus5 et ainsi de suite. Le coefficient d'information mutuelle est calculé avec une fenêtre de 1 024 octets.

Dans notre expérience, nous nous sommes restreints à cent entrées lexicales choisies dans une liste des mots les plus fréquents apparaissant dans le corpus japonais. Les cent mots que nous avons choisis appartiennent au sous-ensemble des mots les plus fréquents obéissant aux conditions suivantes :

- mots possédant au moins deux kanjis ;
- mots apparaissant avec une grande fréquence dans corpus2.

En pratique, bien que les dictionnaires bilingues entre le japonais et l'anglais soient réputés être de bonne qualité, ils ne sont pas exempts d'erreurs. Cependant, nous avons estimé cette qualité comme suffisante pour servir de témoin. Donc, pour mettre en évidence les effets de notre raffinement, les équivalents dans le dictionnaire japonais-anglais des cent mots choisis ont été délibérément dégradés, en supprimant au hasard une bonne moitié, et en ajoutant des traductions inappropriées. Par exemple, *ken-kyuu* était traduit à l'origine par `research`, `work`, `study`, `investigation`, `inquiry`, `examination`, ce qui correspond à un éventail tout à fait honorable d'équivalents. `Study`, `work`, `examination` ont été volontairement remplacés par `tweed` (tweed), `torrid` (torride) et `stover` (étuve, fourneau ou poêle, mais avec une erreur).

5. Analyse de l'expérience et de ses résultats

5.1. L'exemple de *ken-kyuu*

Le nombre de mots japonais qui étaient reliés à *ken-kyuu*, dans nos corpus est de 78. Pour chacun de ces mots, nous avons trouvé de 3 à 7 équivalents en anglais, dans les dictionnaires considérés. En pratique, environ 400 mots anglais étaient concernés par l'expérience. Tous les équivalents qui ont obtenu un score supérieur à 0,01 pour leur poids de correspondance après raffinement utilisant le corpus6 (la totalité des données) sont les suivants :

```
inquiry(0,24), research(0,12), work(0,10), examination(0,075),
exploration(0,074), study(0,034), investigation(0,030), line(0,029),
specialty(0,017), experiment(0,017), question(0,016), test(0,014),
wicker(0,014), labour(0,013), fabrication(0,011), machinery(0,011),
trial(0,011), snapshot(0,011), wonder(0,010), trade(0,010)
```

Remarquons donc que dans les textes considérés, c'est la notion d'enquête (`inquiry`) qui apparaît comme la plus fréquente pour traduire notre terme original de *ken-kyuu*, ce qui est normal pour un corpus journalistique (et qui explique donc le report de la

spécificité du corpus dans le dictionnaire raffiné). Il est ensuite suivi de « recherche » (`research`) pour indiquer la recherche scientifique, lequel est renforcé par le terme suivant immédiatement : « travail » (`work`). Remarquons aussi que les termes `study` (étude), `examination` (examen) et `investigation` (investigation) arrivent en bon rang. Le bruit est en revanche représenté par des termes tels que `line` (ligne), `wicker` (osier), `trade` (profession, commerce), et `snapshot` (instantané) quoique ce dernier terme puisse être relié à la notion dominante d'enquête (policière, juridique,...).

corpus	les 7 mots les plus forts
corpus6	`inquiry research work examination exploration study investigation`
corpus5	`research inquiry work exploration examination line investigation`
corpus4	`inquiry work exploration examination research study line`
corpus3	`inquiry work exploration examination study investigation line`
corpus2	`inquiry examination research work line study question`
corpus1	`examination research line work study inquiry investigation`
morph + syn	`research inquiry line work examination preview study`

TABLEAU 3 Les sept meilleurs équivalents de *ken-kyuu*.

L'évolution incrémentale du résultat après usage de chaque corpus est fournie dans le tableau 3 pour les sept équivalents anglais de *ken-kyuu* les plus importants en terme de poids relatif. Des équivalents assez peu adéquats tels que `line` (ligne) et `preview` (avant-première, annonce) sont apparus dans le champ des sept meilleurs mots après application de MCM_s (matrice des informations d'origine statique, notée « morph + syn » dans le tableau), mais leurs poids tendent à diminuer au fur et à mesure que les corpus considérés s'élargissent. Cela semble indiquer qu'au moins sur les premiers pas de l'algorithme, la qualité de la distribution est proportionnelle à la taille du corpus. En fait on s'aperçoit que les équivalents se stabilisent à partir du corpus4, mais c'est leur ordre qui change. On s'aperçoit aussi que le terme parasite `line` disparaît au niveau du corpus6, qui ne retient plus que des équivalents appropriés.

5.2. Quelques résultats statistiques concernant les 100 mots les plus fréquents

Nous définissons une proportion commune entre deux dictionnaires comme étant le

nombre total d'équivalents (toutes entrées lexicales confondues) divisé par le nombre total d'entrées lexicales communes. Nous notons cette proportion EF (pour *Equivalence Fraction*). Pour chaque mot en LS les équivalents qui ont des poids supérieurs à 0,001 sont conservés, et pour cet ensemble d'équivalents, nous calculons le rappel de la manière suivante : **rappel** est l'EF du dictionnaire dégradé en entrée (noté dictionnaire1) et du dictionnaire témoin japonais-anglais qui sert de référence, noté dictionnaire2. On remarquera qu'à l'origine il y a exactement 45,5 % d'équivalences communes, telles qu'elles apparaissent dans le tableau 4.

corpus	rappel
corpus6	61,8 %
corpus5	63,1 %
corpus4	62,5 %
corpus3	62,8 %
corpus2	55,1 %
corpus1	60,2 %
morph + syn	53,3 %
dégradé	45,5 %

TABLEAU 4 Valeurs de rappel et incrémentalité

Chaque valeur de rappel de ce tableau est calculée après chaque pas dans la procédure incrémentale. Le dictionnaire qui résulte du pas d'incrémentation est comparé avec le dictionnaire témoin.

Le dictionnaire final, résultant de l'application de la procédure incrémentale avec corpus6, fournit un rappel de 61,8 %, ce qui veut dire que le dictionnaire dégradé par nous a évolué depuis une couverture de 45,5 % des équivalents, jusqu'à couvrir environ les deux tiers des équivalents du dictionnaire de référence.

Ainsi, la valeur ajoutée par notre algorithme en terme de rappel seulement est de 16,3 % avec un corpus de 33 méga-octets, qui est de taille modeste lorsque l'on considère des données écrites en kanji, et qui ne provient que de journaux. Cette valeur n'est pas obligatoirement faible dans la mesure où, d'une part effectivement, le corpus n'est pas si grand, et d'autre part, l'usage courant de la langue, tel qu'on peut le voir dans les journaux, n'utilise pas forcément la totalité des associations entre un mot et ses similaires, loin s'en faut. Il existe des synonymes très techniques, et on ne les trouvera pas forcément dans les journaux alors qu'on peut les trouver dans le dictionnaire. Remarquons par ailleurs que pour le mot *ken-kuyu*, l'algorithme a restitué la totalité des équivalents cités : cela signifie que pour des mots de ce type, les corpus journalistiques fournissent une assez bonne couverture en terme d'associations sémantiques, alors que cela peut ne pas être le cas pour d'autres termes, mêmes lorsqu'ils font partie des cents mots les plus employés. On s'aperçoit donc très rapidement que ce sont surtout les résultats qualitatifs de notre algorithme qui sont les plus importants. Outre

ce que nous venons de dire, nous avons obtenu de nouveaux équivalents qui n'étaient pas enregistrés dans le dictionnaire de référence anglais-japonais et qui sont montrés en annexe de cette contribution. Cela signifie notamment que nous sommes en mesure d'enrichir des bases de connaissances lexicales bilingues dynamiquement, avec des associations qui ne sont pas obligatoirement présentes dans des dictionnaires d'usage très général.

A	B	pourcentage
Dégradé	syn + morph	56,1 %
syn + morph	corpus1	94,1 %
corpus1	corpus2	67,2 %
corpus2	corpus3	99,7 %
corpus3	corpus4	91,9 %
corpus4	corpus5	88,4 %
corpus5	corpus6	88,4 %

TABLEAU 5 Convergence de la méthode

Remarquons que la fonction de rappel n'est pas monotone. Ainsi, les valeurs obtenues pour les corpus 3, 4 et 5 sont meilleures qu'avec le dernier corpus, qui est aussi le plus grand en volume. Ce peut être un effet de seuil : à partir d'une certaine taille, plus un corpus est grand, plus certains mots peuvent s'y éparpiller, car les corpus ne sont pas homogènes. Parmi les cent mots sélectionnés, certains ont obtenu une fréquence d'occurrence inférieure à 0,01.

On peut aussi remarquer que le meilleur pourcentage de rappel est fourni au niveau du corpus5. C'est ce qui nous a amenées à nous poser la question de la convergence éventuelle de la méthode. Pour estimer cette convergence, le tableau 5 montre la EF entre le dictionnaire obtenu après l'étape n (appelé A) de l'algorithme et le dictionnaire résultant de l'étape $n + 1$ (appelé dictionnaire B). Les poids correspondants aux critères statiques (score morphologique et de synonymie) modifient l'état du dictionnaire, qui semble alors acquérir une quantité appréciable d'information. En fait, le tableau 4 montre que déjà près de la moitié du gain quantitatif de l'algorithme est réalisé au niveau de l'adjonction de l'information statique. Le corpus commence à influencer le résultat à partir du corpus2. Les effets des corpus 3 et 4 demeurent faibles, alors qu'ils se potentialisent au niveau du corpus5 tout en se stabilisant. Il semble que, de manière globale, 10 % environ des équivalents sont modifiés par le corpus, même lorsque l'on considère le plus important de tous, c'est-à-dire le corpus6. Il est raisonnable de penser que la méthode se stabilise au niveau du corpus5, et que le dictionnaire obtenu après cette étape est le meilleur possible, ce qui conduit à arrêter le processus à ce niveau et à considérer un gain quantitatif de 18 % avec l'algorithme, gain dont on est sûr de la qualité.

5.3. Conclusion sur les résultats

Pour résumer les propos précédents, nous pouvons dire que l'algorithme présenté a permis de concrétiser les points suivants :

- un dictionnaire dégradé (ou de mauvaise qualité) peut gagner numériquement 20 % de bonnes équivalences en plus par la méthode incrémentale ;
- le meilleur score de l'algorithme n'apparaît pas au niveau du corpus le plus grand, mais dans un corpus de taille suffisante assurant la meilleure distribution des occurrences ;
- la fonction de rappel n'est pas monotone, mais semble se stabiliser, ce qui est quand même un indice de sa fiabilité ;
- l'information statique (c'est-à-dire les connaissances obtenues à partir de dictionnaires et de lexiques monolingues en LS) demeure fondamentale, dans la mesure où elle fournit à elle seule la moitié du gain quantitatif et qualitatif de l'algorithme. On ne peut donc pas en faire l'économie, et se contenter de raffiner des dictionnaires exclusivement à partir de corpus ;
- néanmoins, l'information dynamique extraite de morceaux de discours garde tout son intérêt puisqu'elle complète judicieusement cet ensemble d'informations statiques, en fournissant l'autre moitié du gain de raffinement, conséquent à l'application de l'algorithme sur le dictionnaire ;
- la méthode ainsi décrite dote le dictionnaire résultant d'équivalents qui n'étaient pas présents dans le dictionnaire de référence.

Il ne faut cependant pas perdre de vue que notre algorithme possède des limites et nécessite des améliorations, notamment sur les points suivants.

- Le premier point discutable est le choix de la formule de calcul des poids. Ici, les entrées possèdent systématiquement un poids unitaire qui se divise sur les différents arcs afférents : c'est donc une méthode à somme constante. La valeur maximale de 1 est ajoutée (par addition avec la matrice unitaire) pour ne pas avoir d'explosion combinatoire. L'inconvénient d'un tel calcul est que l'on perd la symétrie entre deux entrées qui sont en correspondance. Donc LS et LC ne sont pas interchangeables. Ainsi, dans une situation où le japonais est LS et l'anglais est LC et que l'on raffine le dictionnaire dans le sens japonais-anglais, on ne peut pas simultanément le raffiner dans l'autre sens, en utilisant un corpus en langue anglaise. De plus, les entrées varient avec le nombre de leurs équivalents, dès lors, on ne peut pas associer de prime abord le même poids initial de 1 à toutes les entrées. En pratique, il faudrait pondérer une entrée par rapport à l'ensemble des entrées du dictionnaire.
- Le second point concerne le sort des équivalents dont les scores sont bas. L'algorithme rejette ces équivalents. Cependant, certains peuvent être utiles parce que leur association en contexte avec les entrées concernées indique qu'il existe une relation sémantique d'association, qui n'est pas forcément une relation de synonymie, et qui peut intervenir dans une représentation sémantique conceptuelle de ces mots. On peut supposer que les équivalents dont les poids sont les plus forts sont en relation forte avec l'entrée, que ce soit une relation d'hyperonymie-hyponymie, ou en relation de synonymie. On peut aussi considérer que les équivalents à poids plus faibles ne sont pas absents de l'attribution du sens à un mot, et qu'ils seront alors utiles pour compléter d'une façon pertinente une résolution

de polysémie en contexte. Mais cette réflexion concernant la possibilité de bâtir une véritable représentation conceptuelle à partir d'aspects quantitatifs sur des données fait l'objet d'un autre travail que nous avons entrepris de réaliser pour l'amélioration de dictionnaires à partir de réseaux sémantiques bilingues.

6. Conclusion générale et perspectives

Dans cette contribution, nous avons proposé une méthode de raffinement d'un dictionnaire bilingue de qualité médiocre, à partir de données monolingues fournies dans la langue source. Cette méthode est dite incrémentale, en ce sens qu'elle définit récursivement des matrices-dictionnaires sur des pondérations de termes cooccurrants dans des corpus de la langue source, dont la taille varie à chaque pas de la méthode. Nous avons mis en commun des données aussi bien statiques (lexiques monolingues de la langue source) que dynamiques (associations en discours de lexies sur lesquelles des hypothèses de correspondance sémantique sont faites).

De manière formelle, l'algorithme est centré autour d'une multiplication de deux matrices correspondant l'une à un dictionnaire monolingue, enrichi par des informations morphosémantiques issues d'un lexique, et dynamiquement augmentées par un traitement de corpus, et l'autre à un dictionnaire bilingue dénotant les relations entre les lexies en langue source et leurs équivalents en langue cible.

L'algorithme a été testé avec un dictionnaire japonais-anglais dégradé par suppressions d'une moitié des équivalents et par adjonction de correspondances erronées. Nous avons aussi utilisé un corpus de 33 méga-octets des archives du journal *ASAHI*, et nous avons sélectionné une liste de cent mots qui ont servi à mener l'expérience. Les résultats montrent que l'algorithme a pu restituer la moitié des équivalents manquants, et a introduit de nouveaux équivalents pertinents (voir l'annexe).

Trois voies de recherche ultérieures peuvent être envisagées. Premièrement, rendre symétrique l'algorithme de raffinement qui comme nous l'avons mentionné, donne des résultats dépendant du rôle de chaque langue considérée, ce qui permettra d'utiliser un corpus monolingue dans chacune des deux langues, indifféremment. Deuxièmement, vérifier l'impact des caractéristiques du corpus (taille et nature) pour bien maîtriser les propriétés de l'incrémentalité. Troisièmement, tester l'algorithme avec des corpus typés tels que des corpus scientifiques ou littéraires pour produire des correspondances bilingues plus spécifiques. De toutes façons, quelle que soit la voie qui sera explorée en premier, nous fondons de grands espoirs sur une amélioration des processus automatiques d'extraction de correspondances bilingues. Cela nous mènera assez rapidement vers notre objectif final, qui est la réalisation d'un bon dictionnaire électronique japonais-français et français-japonais, dont le contenu est actualisé par les corpus de discours contemporains.

Remerciements

Les auteurs remercient le Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI) du CNRS, et son directeur Joseph Mariani, pour la mise à disposition de l'environnement de test, l'accueil et les conseils. Le Dr K. Umemura de

Nippon Telegraph and Telephone a gracieusement fourni les données électroniques et nous lui exprimons notre gratitude. Nous remercions de même Dr S. Hayamizu pour ses précieux conseils et ses commentaires intéressants. Le Dr Utsuro de AIST a eu la bonté de nous offrir l'accès à l'analyseur morphologique du japonais JUMAN.

Annexe

japonais	équivalents en anglais
kei-kaku	aim intent meaning purpose idea predestination intention design sense bet schedule message program
jou-hou	communication correspondence mass_communication transmission liaison intercourse junction news telegraph intelligence contact epistle missive letter chou-sa examination exploration test trial research inquiry checkup proof interrogatory audit probation search check hearing medical
hen-ka	mutation alteration demolition cataclysm diversity shift transformation metamorphosis defeat warp transmutation distance switch wreck swing transfiguration vicissitude shuffle
i-ken	apprehension attitude posture understanding judgement position anxiety idea capture bearing attest antenna setup set stance amyloid concern collar assessment unrest care fear view opinion verdict solicitude worry capacity
kou-zou	system scheme dot acer assemble making constitution taylor composition texture frame stout tow structure construction tidal conformation make-up
sho-ri	treatment disposal proceeding cure disposition remedy arrangement
jo-sei	feminine female womankind hen she bitch cow woman daughter
nin-ki	popularity star epidemic boom fashion report notoriety repute reputation currency name child whisper mode