Le traitement de la phraséologie dans DEFI

Archibald Michiels
Département d'anglais - Université de Liège (Belgique)

DEFI is a prototype computer tool aimed at ranking (from most to least relevant) the French translations of an English lexical item in context. This paper deals with the strategies used by DEFI to recognize multi-word units (mwus) in running text. Any lexical unit included in the lexical database used in the project (a merge of the Oxford/Hachette and Robert/Collins English-to-French dictionaries) and longer than a single word is submitted to a surface parser, and the same process is applied to the user's text. A program written in Prolog assesses the quality of the match between the parsed user's text and candidate mwus retrieved from the project's lexical database. The matcher is able to account for some of the distortions undergone by the mwu, e.g. movement of a constituent as a result of relativization or passivization.

DEFI est un outil d'aide à la lecture de textes anglais pour lecteurs francophones. Il agit comme filtre sur un dictionnaire bilingue anglais-français pour ne retenir que les acceptions qui conviennent au contexte, et en présenter les traductions au lecteur dans un ordre de pertinence décroissante. Si l'item pour lequel le lecteur a demandé de l'aide est un élément d'une unité phraséologique, le système propose cette unité et la traduction la mieux ajustée au contexte.

De toute évidence, le paragraphe précédent décrit une situation optimale. DEFI ne dispose que des informations qui sont contenues dans sa base de ressources lexicales et qui sont suffisamment formalisées pour qu'il puisse en faire usage, et son analyse du texte source, basée sur un parseur de surface (*Lingsoft engcg parser*), est assez rudimentaire et parfois erronée. DEFI est un outil applicable à tout texte rédigé en anglais: le revers de cette médaille de généralité est bien sûr la superficialité de l'analyse et, partant, dans bien des cas, l'impossibilité d'effectuer un choix fondé entre les diverses traductions proposées par le dictionnaire.

Le lecteur trouvera sur le site Web de DEFI (http://engdep1.philo.ulg.ac.be/michiels/efdefi.htm) une vingtaine de contributions qui lui permettront de se faire une idée assez nette du fonctionnement de DEFI, et aussi d'évaluer le système sur la base des résultats qu'il offre pour un millier de phrases soumises à l'analyse (http://engdep1.philo.ulg.ac.be/michiels/cobres.htm). Les contributions sont en anglais, à l'exception de http://engdep1.philo.ulg.ac.be/michiels/defifr.htm. Pour ce qui est des publications papier, on se reportera à Michiels 1998, Michiels & Dufour 1998 et Michiels 2000.

Je voudrais ici concentrer l'attention sur le traitement des unités phraséologiques dans DEFI. Je n'insisterai pas sur la prise de conscience, somme toute assez récente, de l'importance de la phraséologie dans les processus de compréhension et de traduction de texte. Je voudrais seulement souligner

que le dictionnaire bilingue est bien souvent une source plus riche que le monolingue (à granularité égale, s'entend) pour les unités phraséologiques, car le caractère semi-compositionnel de bon nombre d'entre elles permet au lexicographe monolingue d'en omettre la mention, alors que le lexicographe bilingue, qui doit traduire, ne peut les négliger que si leur traduction résulte de la concaténation de traductions proposées pour leurs éléments constitutifs sous leurs propres entrées, cas somme toute assez rare. Ainsi LDOCE ne se fera pas scrupule d'utiliser *take action* dans la définition même de *act*, alors que RC et OH rendront compte de *take action* comme unité phraséologique, et y associeront une traduction qui ne résulte pas de la concaténation d'une traduction de *take* avec une traduction de *action* ('agir, prendre des mesures'). On peut aussi se convaincre de la richesse phraséologique du bilingue en comparant le nombre d'unités phraséologiques répertoriées sv *picture* dans OH et RC d'une part, et la relative indigence à cet égard de monolingues de granularité comparable (LDOCE, COBUILD et CIDE).

On objectera toutefois que le bilingue, encore plus que le monolingue, tend à mélanger phraséologie et exemplification. Si le bilingue semble être plus riche en phraséologie, c'est en partie parce qu'il ne fait pas le départ entre ce qui est vraiment unité phraséologique (compositionnalité et variabilité restreintes) et ce qui est illustration d'une acception donnée dans un syntagme ou une phrase d'exemple. On conviendra que la distinction est particulièrement difficile à établir quand l'unité phraséologique doit être accompagnée d'une traduction. Il est plus aisé de la présenter dans un syntagme plus large, ou, plus souvent encore, toute une phrase, et d'offrir une traduction de cet ensemble (*to miss someone*: 'regretter l'absence de quelqu'un' versus *I miss you*: 'tu me manques').

Dans DEFI nous avons choisi de considérer comme unité phraséologique tout ce qui dépasse la taille du mot. Tout ensemble de plusieurs mots est ainsi *ipso facto* versé dans la base de données d'unités phraséologiques. Se pose alors le problème de l'accès. Il est illusoire de croire que le lecteur qui demande de l'aide en pointant un élément d'une unité phraséologique a reconnu dans le texte cette unité et de surcroît peut déterminer avec exactitude le mot vedette dans l'entrée duquel le lexicographe a décidé de lui faire place. Il faut considérer au contraire que ce lecteur n'a peut-être même pas conscience de pointer vers un élément d'un ensemble lexicologique plus large. On s'assurera donc que tout élément (à l'exception des mots outils) d'une unité phraséologique puisse servir de point d'accès pour saisir cette unité et la soumettre au système pour qu'il calcule la qualité de l'appariement unité phraséologique – texte de l'utilisateur.

Pour ce faire, nous avons procédé comme suit:

1. Fusion de nos deux dictionnaires bilingues anglais-français, RC et OH. Cette fusion a été accomplie de manière très conservatrice. On a plutôt visé à ne pas perdre d'informations pertinentes qu'à éliminer la redondance (cf. http://engdep1.philo.ulg.ac.be/michiels/merge.htm).
2. Extraction des unités dépassant la taille du mot (nous les appelle-

rons lexies), qu'il s'agisse de composés nominaux (*dog biscuit*, *dog breeder*), d'expressions idiomatiques plus ou moins figées (*it's a case of the tail wagging the dog*, *let sleeping dogs lie*) ou de phrases d'exemple, illustrant une acception ou une unité phraséologique (*let the dog in and give it a drink*, *misfortune dogs his footsteps*)

3. Production de listes d'accès, c'est-à-dire d'ensembles de mots par lesquels une lexie peut être 'saisie'. La figure 1 donne l'ensemble des listes d'accès associé à l'item *bell*.
4. Parsage de toutes les lexies à l'aide de **engcg** de Lingsoft et calcul de structures (groupes nominaux, hypothèse structurale pour toute la lexie), de relations (relations syntaxiques nécessaires à l'identification des collocats) et propriétés (voix, polarité). Les lexies du dictionnaire et le texte de l'utilisateur sont soumis à une même analyse pour pouvoir exploiter à fond les similitudes de tout ordre qui réduisent la distance dictionnaire-texte ou, sous un angle plus positif, améliorent la qualité de l'appariement entre les deux. Les figures 2 et 3 montrent ce parallélisme dans le traitement de la lexie et du texte de l'utilisateur.

```
/* la liste d'accès pour bell donne toutes les clauses d'accès des lexies comportant le
lemme bell */
inh('bell',
   [
     [[p_h,answer,answer],[p,bell,bell]], /* p_h désigne le mot vedette, ici answer */
     [[p_h,bell,bell],[p,bottomed,bottom],[p,trousers,trouser]],
     [[p_h,bell,bell],[p,bottoms,bottom]],
     [[p_h,bell,bell],[p,cat,cat]],
     [[p_h,bell,bell],[p,door,door]],
     [[p_h,bell,bell],[p,first,first],[p,t_prep,for],[p,mass,mass],[p,ringing,ring],[p,t_be,be]],
     /*les entrées pour les mots outils donnent la catégorie en lieu et place de la variante
       morphologique:
       for est donné comme préposition, be comme verbe copule */
     [[p_h,bell,bell],[p,give,give]],
     [[p_h,bell,bell],[p,great,great]],
     [[p_h,bell,bell],[p,name,name],[p,rings,ring]],  /* that name rings a bell */
     [[p_h,bell,bell],[p,number,number],[p,rings,ring]],
     [[p_h,bell,bell],[p,pull,pull]],
     [[p_h,bell,bell],[p,push,push]],
     [[p_h,bell,bell],[p,ring,ring]],
     [[p_h,bell,bell],[p,ringer,ringer]],
     [[p_h,bell,bell],[p,rope,rope]],
     [[p_h,bell,bell],[p,shaped,shape]],
     [[p_h,bell,bell],[p,t_be,be],[p,there,there]],
     [[p_h,bell,bell],[p,t_mod,can],[p,hear,hear]],
     [[p_h,bell,bell],[p,tent,tent]],
     [[p_h,bell,bell],[p,tower,tower]],
     [[p_h,bells,bell],[p,ring,ring]],
     [[p_h,bells,bell],[p,sound,sound]],
     [[p_h,bells,bell]]
   ]
 ).
```

Figure 1: Exemple de liste d'accès

```
dic(
% numéro d'identification :
251141,

% liste d'accès à la lexie :
[[p_h,bell,bell],[p,name,name],[p,rings,ring]],

% clause de type w correspondant au premier mot, that :
[w(0,1,text('that',l),
      lem('that',l),morph([m(pos,det,2),m(type,central,0),
      m(type,dem,2),m(num,sg,2)]),
      syn([s(type,dn,0,r)])),

% clause de type w pour le mot name :
w(1,2,text('name',l),
      lem('name',l),morph([m(pos,n,5),
      m(case,nom,0),m(num,sg,2)]),
      syn([s(func,subj,5,_)])),

w(2,3,text('rings',l),
      lem('ring',l),morph([m(pos,v,5),m(tense,pres,1),
      m(num,sg3,1),m(type,finite,2)]),
      syn[s(type,main,3,f)])),

w(3,4,text('a',l),
      lem('a',l),morph([m(type,indef,3),
      m(pos,det,2),m(type,central,0),m(type,art,3),m(num,sg,2)]),
      syn([s(type,dn,0,r)])),

w(4,5,text('bell',l), % forme de la variante morphologique
      lem('bell',l), % lemmatization
      morph([m(pos,n,5),m(case,nom,0),m(num,sg,2)]), % traits morphosyntaxiques
      syn([s(func,obj,5,_)])),

punct(5,6,unkn)], % ponctuation à la fin de la liste de mots

[np(0,2,c(1,2)),np(3,5,c(4,5))], % liste des groupes nominaux (np's)

[csubj('name','ring'), % relations syntaxiques
cdobj('bell','ring')],

neg(0), % polarité
passive(0), % diathèse
s, % hypothèse structurale
le($that name rings a bell$), % lemme
sc([nil]), % collocats sujets
oc([nil]), %  collocats objets
env([nil]), % environnement syntaxique
pat(nil), % modèle pour verbes prépositionnels
lab([nil]), % étiquettes
sst([nil]), % traits stylistiques
gt(nil), % référence croisée
tr($ce nom me dit quelque chose$), % traduction
rat(1,1), % fréquence de la traduction
gl(nil), % glose
ohef). % origine
```

Figure 2: Clause Prolog extraite du dictionnaire DEFI *des lexies: (lexie 'That name rings a bell' extraite de* OH*)*

```
txt(m, % mode –m : appel à l'appariement d'une lexie
['bell'], % mot sélectionné
[w(0,1, % premier item de la liste de mots: does
      text('does',u), % variante morphologique
      lem('do',u), % lemme
      morph([m(pos,v,5),m(tense,pres,1),m(num,sg3,1),m(type,finite,2)]),
      syn([s(type,aux,3,f)])),
w(1,2, % it
      text('it',l),
      lem('it',l),
      morph([m(type,nonmod,0),m(pos,pron,2),m(case,nom,0),m(num,sg3,1),
      m(func,subj,3)]),
      syn([s(func,subj,5,_)])),
w(2,3, % ring : première lemmatisation : en tant que nom
      text('ring',l),
      lem('ring',l),
      morph([m(pos,n,5),m(case,nom,0),m(num,sg,2)]),
      syn([s(func,subj,5,_),s(func,obj,5,_),s(func,i_obj,5,_)])),
w(2,3, % ring: deuxième lemmatisation : en tant que verbe
      text('ring',l),
      lem('ring',l),
      morph([m(pos,v,5),m(mood,inf,2)]),
      syn([s(type,main,3,nf)])),
w(3,4, % a
      text('a',l),
      lem('a',l),
      morph([m(type,indef,3),m(pos,det,2),m(type,central,0),
      m(type,art,3),m(num,sg,2)]),
      syn([s(type,dn,0,r)])),
w(4,5, % bell
      text('bell',l),
      lem('bell',l),
      morph([m(pos,n,5),m(case,nom,0),m(num,sg,2)]),
      syn([s(func,subj,5,_),s(func,obj,5,_)])),
punct(5,6,comma),
w(6,7, % Malcolm
      text('malcolm',u),
      lem('malcolm',u),
      morph([m(type,proper,3),m(pos,n,5),m(case,nom,0),m(num,sg,2)]),
      syn([s(func,obj,5,_),s(func,app,2,_)])),
punct(7,8,qmark)],


[np(1,2,c(1,2)),np(2,3,c(2,3)),np(3,5,c(4,5)),np(6,7,c(6,7)),
np(1,3,c(1,2)),np(2,5,c(2,3))],
% liste de gn


[csubj('it', 'ring'),csubj('ring', 'ring'),cdobj('ring', 'ring'),
ciobj('ring', 'ring')],
% relations syntaxiques
% remarquez qu'elles sont erronées, l'erreur provenant de la lemmatisation de ring
comme nom
neg(0),
% polarité


 passive(0),
% diathèse
```

```
 s,
% hypothèse structurale

 $Does it ring a bell, Malcolm ?$).
% texte utilisateur
```

Figure 3: Clause Prolog correspondant au texte utilisateur
'Does it ring a bell, Malcolm?' (John Le Carré, 'The Little Drummer Girl')

Si on se place à un niveau d'abstraction suffisamment élevé, on comprendra que la différence entre acception à déterminer par le contexte et élément d'unité phraséologique ne peut être qu'une différence de degré, à savoir le degré de lexicalisation de l'environnement de l'item. Il est certain que si cet environnement peut se définir au niveau de la forme textuelle (*go to the **dogs***) ou du lexème (*top **dog***) et pas seulement de la catégorie thésaurique ou d'un groupement sémique (*to keep a **dog** on a leash*), on parlera d'unité phraséologique. Il nous semble donc plus prudent

1. de ne rejeter hors du domaine de la phraséologie aucun ensemble lexicographique de plus d'un mot attesté dans notre dictionnaire bilingue
2. d'adopter un système qui mesure la distance qui sépare l'unité phraséologique telle que répertoriée dans le dictionnaire et telle que potentiellement instanciée par le texte de l'utilisateur.

Pour l'accomplissement de cette deuxième tâche, deux méthodologies s'affrontent:

A. Développer pour chaque unité phraséologique *sensu stricto* une grammaire locale qui stipule de manière tout à fait univoque les transformations que cette unité peut subir tout en conservant sa nature phraséologique. Ces transformations sont syntaxiques (passivisation, relativisation, pluralisation) mais aussi lexicales (remplacement d'un élément de l'unité par un synonyme, insertion de matériau étranger en position de modificateur d'un des éléments, etc.). La grammaire locale est transformée en un transducteur à nombre fini d'états qui réussit ou échoue face au couple lexie/texte utilisateur. Les exemples donnés par le lexicographe sont ramenés si possible au squelette phraséologique qui les sous-tend. Là où cette réduction n'est pas possible, ils sont négligés. Cette première approche caractérise le projet **Locolex** de Rank Xerox (cf. Bauer et al. 1995; Breidt et al. 1996; Segond & Breidt 1996).
B. Soumettre le texte de l'utilisateur et la lexie candidate à l'appariement (lexie au sens très large de toute unité plus grande que le mot ayant fait l'objet d'un traitement lexicographique, y compris l'exemplification) au même processus d'analyse et sur base des résultats de ce processus mesurer la distance qui les sépare. Cette deuxième approche est celle de DEFI.

Nous pouvons apporter les arguments suivants en faveur de notre approche:

1. Les dictionnaires existants, et notamment RC et OH qui sont à la base de notre dictionnaire bilingue, ne considèrent pas la délimitation du domaine phraséologique comme une priorité. En conséquence, comme on l'a dit plus haut, ils n'hésitent pas à illustrer une unité phraséologique par une phrase entière plutôt que d'en présenter le squelette. L'élément déterminant semble être la nécessité de fournir une traduction qui soit naturelle et conforme au génie de la langue cible. Il n'est donc pas possible de s'en tenir aux unités phraséologiques répertoriées comme telles. Même lorsqu'il est possible de dégager une unité phraséologique et d'en fournir séparément l'illustration (*to ring a bell* – 'dire quelque chose'; *does that name ring a bell?* – 'ce nom vous dit-il quelque chose?'), le dictionnaire opte souvent pour le maintien de l'illustration et l'abandon du squelette phraséologique. En effet, ce squelette phraséologique aurait dû être accompagné d'une liste de collocats (*name, number*, etc.) afin que dire quelque chose ne soit pas interprété comme verbe de dire. Comme la liste de collocats ne présente pas une unité sémique ou thésaurique, le lexicographe opte pour l'illustration toute seule*: that name rings a bell* – 'ce nom me dit quelque chose'; *that number rings a bell* – 'ce numéro me dit quelque chose').

2. La phraséologie est le domaine par excellence de la créativité lexicale. Les formes les plus figées sont toujours susceptibles d'être manipulées. Il semblerait que la seule condition soit la possibilité de recouvrer l'unité sous-jacente, et donc la confiance que le producteur met dans les capacités du récepteur. Il n'est pas étonnant qu'on ne puisse la cerner à l'aide de décisions méthodologiques qui guideraient le travail du lexicographe. Par exemple, on pourrait penser que *shoot the breeze* ('papoter, bavarder') est parfaitement figé (cf. Weinreich 1980:237-247). On ne s'étonnera toutefois pas de trouver dans le dernier roman de Salman Rushdie, *The Ground Beneath Her Feet* (London: Vintage, 2000, 62): *Other parents… strolled off to take (and also shoot) the breeze.*

3. Même si le texte doit être pris en partie au sens littéral, le lecteur appréciera d'obtenir l'unité phraséologique que le texte a frôlée. La lecture se trouvera enrichie de la reconnaissance du procédé allusif mis en œuvre. Mais on ne peut bien sûr pas dissimuler les erreurs d'appariement derrière un vague procédé d'allusion que le développeur de DEFI serait le seul à reconnaître.

D'autre part, il faut bien admettre que la détermination des grammaires locales, qui joue un rôle clé dans la première approche, n'est pas chose aisée. Car de deux choses l'une: ou bien ces grammaires sont développées par des lin-

guistes qui disposent de toutes les informations dont les lexicographes ont disposé lors de la rédaction du dictionnaire, et l'effort de formalisation a alors des chances d'être payant, mais il est extrêmement coûteux en ressources humaines de haut niveau; ou bien ces grammaires sont déterminées sur base des données lexicographiques répertoriées dans le dictionnaire, par un procédé entièrement automatique. Dans ce deuxième cas, on se retrouve en face d'une variante de l'approche DEFI, le transducteur offrant une méthode pour calculer la qualité de l'appariement texte-lexie.

L'appariement que DEFI doit tenter d'établir n'est donc pas un appariement d'une unité phraséologique avec une de ses manifestations textuelles, mais bien plutôt l'appariement de deux manifestations textuelles d'une unité phraséologique, l'une plus contrôlée, celle du dictionnaire, l'autre, plus libre, celle du texte de l'utilisateur. Pour apparier *that name rings a bell* (dictionnaire) et *does that ring a bell?* (texte) il faut non seulement un transducteur capable d'apparier les variantes morphologiques et les lexèmes, mais il faut encore savoir quels éléments peuvent être 'sautés' (*that name*) et exploiter à fond tous les points de similitude entre les deux manifestations textuelles: *bell* est des deux côtés le lexème *bell*, il est des deux côtés au singulier, et il est des deux côtés en relation syntaxique d'objet par rapport au même prédicat *ring*. DEFI profite à fond du mécanisme de remontée et du caractère non-déterministe de Prolog pour débusquer toutes les possibilités d'appariement, leur donner un poids, et les présenter à l'utilisateur dans un ordre de pertinence décroissante.

L'algorithme d'appariement de DEFI comporte les étapes suivantes:

1. vérification d'un seuil de correspondance lexicale entre la lexie (sélectionnée sur base de la présence dans sa liste d'accès du mot pour lequel le lecteur a demandé de l'aide) et le texte de l'utilisateur. L'appariement lexie/texte n'est tenté que si ce seuil est atteint.
2. contrôle de correspondance de voix et de polarité. La voix passive et la polarité négative sont considérées comme **marquées** et doivent correspondre de lexie à texte. Par contre, voix active et polarité positive dans la lexie peuvent correspondre à voix passive et polarité négative dans le texte de l'utilisateur.
3. parcours de la lexie et appariement de chaque élément au texte de l'utilisateur. Peuvent se présenter les cas suivants:
   - correspondance de la forme et du lemme: *dogs* vs *dogs* en tant que nom pluriel; *dogs* vs *dogs* en tant que troisième personne du singulier, indicatif présent
   - correspondance du lemme: *dogs* vs *dog* en tant que nom; *dogs* vs *dog* en tant que verbe
   - appartenance au même ensemble de verbes supports (intr: *become*, *come*, *get*, *go*, *look*, *seem*, *sound*, *turn*; tr: *do*, *get*, *give*, *have*, *lay*, *make*, *put*, *set*, *take*)
   - correspondance entre proforme dans la lexie et structure dans le texte (*something*: gn, *do so*: gv, etc.)

- correspondance zéro (*silent move*) dans la lexie. Il s'agit surtout d'éléments en tête de lexie, qui servent à donner la structure de syntagme ou de phrase: *to, to be, to have, pron+be, there+be, pron+have,* etc. Mais on a été amené à permettre le parcours silencieux du groupe nominal, pour rendre compte des appariements du type *that name rings a bell* / *the title didn't ring a bell*. On notera que la négation peut faire l'objet d'un *silent move* puisque le contrôle de la polarité s'accomplit à un niveau supérieur (ce qui permet de prendre en compte des transformations qui entraînent le mouvement de la négation hors de la proposition sur laquelle elle porte).
- *silent move* dans le texte de l'utilisateur. Cette option est nécessaire pour permettre de traverser les éléments enchâssés dans la réalisation textuelle de la lexie (modificateurs, etc.)

Ce parcours pourrait très bien s'exprimer via un transducteur. Dans les cas de correspondance lexicale (les deux premiers cas) DEFI y ajoute un calcul de la correspondance des traits morphosyntaxiques calculés par le parseur sur les deux textes, le texte de la lexie et le texte de l'utilisateur. Chaque type de correspondance fournit un poids, et c'est l'accumulation de ces poids qui révèle le degré de qualité de l'appariement. DEFI affecte ce poids général d'un coefficient de pondération qui diffère selon la structure de la lexie, unité phraséologique proprement dite (gn, gv, gp) vs toute une phrase (présumée être un exemple).

DEFI doit prendre en compte les distorsions syntaxiques auxquelles une lexie peut être soumise sans pour autant perdre son unité phraséologique. Il s'agit avant tout des manipulations syntaxiques qui affectent le groupe nominal et qui se manifestent par un mouvement de ce groupe hors de sa position canonique. Il n'y a pas lieu de débattre ici de la 'réalité' de ce mouvement; la seule chose que DEFI doit être capable de faire, c'est d'associer la manifestation textuelle de la lexie à sa représentation canonique (ou plus souvent, dans notre cas, à une manifestation textuelle plus proche du squelette canonique).

Les 'transformations' qui impliquent un 'mouvement' du gn sont bien connues; nous considérerons les deux plus importantes, à savoir la passivisation et la relativisation. Nous donnons ci-dessous quelques exemples tirés du *British National Corpus* (BNC), et affectant les lexies *bear the brunt of* et *wreak havoc on*:

Pass1. *the real brunt of the war was being borne by the men on the battlefield.*
Pass2. … *despite the havoc wreaked by Hurricane Iniki.*
Rel1. … *but with the havoc it is wreaking on their faces.*
Rel2. *The havoc they wreaked was total.*

Puisque la réalité du mouvement du gn ne nous importe pas, et que nous ne présentons pas une théorie linguistique qui ambitionnerait d'expliquer les phénomènes grammaticaux, nous pouvons nous tourner vers une procédure

d'appariement qui est résolument atypique en ceci qu'elle transporte le mouvement du groupe nominal vers le verbe principal. Nous agissons de la sorte car le verbe principal est un élément unique, facile à repérer, alors que le gn est un syntagme dont le degré de complexité peut être énorme (présence de déterminants, modifications sous forme de groupe prépositionnels et de relatives, compléments de type propositionnel, etc.) et dont le centre n'est pas aisé à déterminer.

Notre procédure est la suivante. On se souvient que l'appariement d'une lexie (telle que représentée dans le dictionnaire) à sa manifestation textuelle dans le texte de l'utilisateur, s'accomplit des deux côtés par la progression d'un pointeur, progression qui est régie par la nature des éléments des deux chaînes et les possibilités d'appariement d'élément de chaîne à chaîne. On se souvient également que la direction de l'appariement est lexie → texte. Si dans le parcours de la lexie nous rencontrons un verbe principal, nous acceptons l'appariement de ce verbe à son correspondant dans le texte, quelle que soit la position de ce dernier dans le texte. Nous exigeons seulement qu'il s'agisse d'un même lexème, et qu'il ait été parsé comme verbe principal à la fois dans la lexie et dans sa manifestation textuelle. Simultanément, nous retenons la position qui suit immédiatement le verbe dans le texte en instanciant une variable Prolog (*Resumption*) que nous passons au prédicat principal, responsable du traitement du mouvement du groupe nominal. Le pointeur sur le texte est ramené en début de chaîne. Quand dans la suite nous tentons d'apparier un quelconque élément de la lexie à son correspondant textuel, nous permettons au prédicat *gothrough*, celui qui fait avancer le pointeur dans le texte sans mouvement concomitant dans la lexie (*silent move*), de placer le pointeur sur la position mémorisée dans la variable *Resumption*, pour autant que cette variable ait été instanciée. Nous pouvons dès lors procéder à l'appariement des autres éléments de la complémentation verbale, ceux qui n'ont pas été affectés par le mouvement du groupe nominal.

Donnons un exemple. Soient à apparier la lexie *wreak havoc on something* et sa manifestation textuelle *the havoc it is wreaking on their faces* (Rel1):

lexie: [0] to [1] wreak [2] havoc [3] on [4] something [5].
texte: [0] the [1] havoc [2] it [3] is [4] wreaking [5] on [6] their [7] faces [8]

1. Le pointeur est positionné en début de lexie et en début de texte (position 0). Le marqueur *to* est traversé dans la lexie sans appariement au texte (*silent move*).
2. Grâce à l'étape préalable de lemmatisation et parsage, *wreak* a été reconnu comme verbe principal, à la fois dans la lexie et dans sa manifestation textuelle (*wreaking*). Le *wreak* de la lexie (entre les positions 1 et 2) peut donc être apparié avec le *wreaking* du texte (positions 4-5), pour autant que la variable *Resumption* soit instanciée à la position de sortie du verbe, à savoir 5.
3. Dans la foulée, le pointeur sur le texte est remis à zéro. L'élément *havoc* de la lexie (2-3) est apparié au *havoc* du texte (1-2), le

déterminant du texte (*the*) étant un élément traversable par le prédicat *gothrough*.

4. Le *on* (3-4) de la lexie est apparié au *on* du texte (5-6), car le prédicat *gothrough* peut déplacer le pointeur sur la position indiquée par la variable *Resumption*, instanciée à 5 à l'étape 2.
5. La pro-forme *something* est appariée au gn *their faces* (6-8).
6. L'appariement réussit puisque le pointeur est en position finale de la lexie; sa pondération est calculée (167 – voir ci-dessous), et, partant, la qualité de l'appariement (il occupe la première position et c'est donc le meilleur – les appariements de qualité inférieure ne sont pas repris ci-dessous).

DEFI produit le résultat suivant:

167 /* pondération de l'appariement /
- 321686, /* numéro d'identification de la lexie /
ohef, /* origine: OH /
[havoc], /* mot sélectionné par l'utilisateur dans le texte */
to wreak havoc on sth, /* lexie ayant fait l'objet de l'appariement/
tr(dévaster qch), / * traduction retenue */
/* suivent les informations sur le mécanisme d'appariement (réservées au déboguage): */
[
m(c1,vac,to,0), /* le *to* est traversé – pondération de zéro */
m(c6bis,dic(wreak),txt(wreaking),morph(0),syn(3),30), /* appariement de *wreak* – pondération de 30 car il a fallu recourir à la lemmatisation pour obtenir l'appariement – la comparaison des traits syntaxiques attribués à *wreak* des deux côtés permet d'augmenter la pondération de 3/
m(c5,dictxt(havoc),morph(10), syn(0),50), /* appariement de *havoc* – pondération maximale de 50 et bonus de 10 pour congruence de traits morphologiques*/
m(c5,dictxt(on),morph(2), syn(5),50), /* appariement de *on* – pondération maximale et boni de 2 et de 5 pour congruence de traits morphologiques et syntaxiques /
m(c13,dic(sth_pro),txt(faces),12), /* appariement de la pro-forme *something* avec le gn dont la tête est *faces* – pondération de 12 */
m(c4,vac,punct,0) /* ponctuation signalant la fin de la lexie – pondération de zéro */
]

On pourrait envisager d'assortir ce traitement du mouvement du groupe nominal d'un contrôle visant à déterminer qu'un mouvement de gn a bien eu lieu, c'est-à-dire qu'il y a bien eu passivisation ou relativisation impliquant le gn en question. Pour ce qui est de la passivisation, DEFI dispose d'un dra-

peau *passive(Valeur)*, qui prend les valeurs *passive(1)* ou *passive(0)*, et qui est calculé sur base des résultats renvoyés par le parseur. Ce drapeau offre une assez bonne fiabilité. Par contre, le parseur ne donne pas d'indications (excepté les positions des éléments, bien entendu – mais c'est précisément ce dont nous envisageons ici de ne pas nous contenter) sur la topicalisation ou la relativisation à relatif zéro, telle que celle exemplifiée dans Rel1 et Rel2 ci-dessus. Il semble plus expédient de renoncer à des contrôles qui s'avèrent de toute façon insuffisants, et de laisser le mécanisme agir librement.

Il est évident qu'on s'expose par là à laisser passer du bruit, c'est à dire à admettre l'appariement de structures où le gn soi-disant 'déplacé' n'est pas à rattacher à la lexie. Mais il convient de se rappeler que DEFI n'est pas un outil destiné à être confronté à des inputs aussi sévèrement agrammaticaux que ceux qui résulteraient d'un non-respect des lois de saturation des arguments. Il faut aussi se souvenir que la procédure mise en œuvre ici par DEFI n'exclut nullement les autres procédures d'appariement dont il dispose – c'est la pondération qui décidera.

Qu'en est-il des autres mouvements syntaxiques qui peuvent affecter des constituants de lexie? On peut faire un sort rapide aux mouvements des adjectifs en indiquant qu'il n'est guère opportun de les laisser agir sur des lexies, car ils sont souvent destructeurs de l'unité phraséologique. Le mouvement qui sort l'adjectif du groupe nominal pour en faire un attribut du sujet ou de l'objet est fréquemment de ce type. Comparez

> *a lucky dog (un veinard) – the dog is lucky to get such a big bone*
> *a fat cat (une huile) – our cat is getting too fat*
> *a poor fish (un pauvre type) – the fish was poor but the meatloaf was first class*

Les groupes prépositionnels méritent plus d'attention. Les mouvements de gp qui font partie intégrante de lexies sont assez peu fréquents, et pourraient être négligés. Exemple: *To put the cat among the pigeons → They were the pigeons the cat had been put among.*

Il n'en va pas de même des gp qui, à l'intérieur de la lexie, ne font qu'indiquer l'existence d'arguments rattachés à la lexie par le biais de la préposition, qui seule fait partie intégrante de la lexie. Ces gp ont, à l'intérieur de la lexie, la forme suivante: préposition + pro-forme de gn, par exemple *for something*, *to somebody*, *with something*, etc.

Considérons la lexie *to give somebody credit for something*. De même que nous disposons d'une procédure d'appariement des pro-formes de gn, *somebody* et *something*, il nous faut prévoir un traitement du mouvement dont peuvent être affectées deux structures, la pro-forme de gn et le groupe prépositionnel tout entier:

- *the action for which he was given credit* (mouvement du gp tout entier, la préposition ayant été *pied-piped*, pour reprendre l'amusante métaphore propagée par la grammaire transformationnelle)
- *the action he was given credit for* (seul le gn *the action* a subi un mouvement)

Le second cas est simple à traiter dans notre optique. Nous admettons un appariement 'à vide' de la pro-forme de gn qui suit une préposition (nous apparions for *sth* avec *for*, pour autant que les deux pointeurs soient correctement positionnés lors de la tentative d'appariement).

Le premier cas est pris en compte par la procédure suivante. DEFI accepte comme appariement du groupe *prep + pro-forme de gn* du côté de la lexie le groupe *prep + relatif ou interrogatif* (*wh-word*: identifié par un trait morphologique dans l'output du parseur) du côté du texte de l'utilisateur, quelle que soit la position de ce groupe dans ce texte. Cet appariement ne déplace pas le pointeur du texte.

Voici le résultat de l'appariement de la lexie *to give sb credit for sth* et du texte *I don't approve of the action he was given credit for*:

148 - 293381,
ohef,
[credit],
to give sb credit for sth,
tr(attribuer à qn le mérite de qch), [
m(c1,vac,to,0),
m(c6bis,dic(give),txt(given),morph(0),syn(3),30),
m(c12,dic(sb),txt(he),12),
m(c5,dictxt(credit),morph(7),syn(5),50),
m(c30,dic(for),txt(for),60),
m(c4,vac,punct,0)]

Et voici les résultats pour le texte *I don't approve the action for which he was given credit*:

138 - 293381,
ohef,
[credit],
to give sb credit for sth,
tr(attribuer à qn le mérite de qch), [
m(c1,vac,to,0),
m(c6bis,dic(give),txt(given),morph(0),syn(3),30),
m(c12,dic(sb),txt(he),12),
m(c5,dictxt(credit),morph(7),syn(5),50),
m(c29,dic(for),txt(for),50),
m(c4,vac,punct,0)]

Dans son évaluation de la qualité de l'appariement, DEFI tient aussi compte de l'ancrage de la lexie dans son contexte, notamment via le calcul de la correspondance entre les collocats spécifiés dans l'entrée lexicale et les items occupant les positions syntaxiques correspondantes dans le texte de l'utilisateur. Ce calcul fait appel à des ressources lexicales monolingues telles que le thésaurus de *Roget* et *WordNet* car les relations thésauriques sont ici primordiales. Les collocats donnés par le

dictionnaire ne doivent en effet pas s'interpréter uniquement comme des mots morphosyntaxiques ou des lemmes, mais aussi comme têtes de catégories thésauriques (*dog* recouvre les mot *dog* et *dogs*, mais aussi *poodle* etc.).

DEFI a deux modes de fonctionnement. Celui qui nous occupe ici est le mode *–m* (*multi-word unit mode*), dans lequel DEFI met en œuvre l'algorithme d'appariement lexie/texte. Dans le mode *–s* (*single-word unit mode*), il tente de déterminer l'acception en contexte d'un lexème. Dans l'état actuel de DEFI, c'est à l'utilisateur qu'il appartient de déterminer le mode de fonctionnement qu'il souhaite. Pour rendre la procédure automatique, il faudrait spécifier un seuil de qualité en-deça duquel les résultats renvoyés en mode *–m* ne peuvent tomber. Le passage au mode *–s* se ferait alors automatiquement. Cette procédure est à la fois délicate à mettre à point et d'un coût computationnel élevé. Mais il est certain que le mode actuel de fonctionnement n'est pas optimal, dans la mesure où, comme nous l'avons fait remarquer plus haut, l'utilisateur ne soupçonne pas toujours que l'élément pour lequel il demande de l'aide est en fait partie d'une lexie.

L'évaluation des résultats de DEFI est particulièrement difficile, notamment pour les raisons évoquées dans http://engdep1.philo.ulg.ac.be/michiels/cobres.htm. Les sources d'erreur sont en effet nombreuses: elles vont de déficiences dans les dictionnaires de base, en passant par les erreurs du parseur, aux problèmes liés à l'algorithme d'appariement et à la pondération des équivalences. Le fichier test *cobres.htm* permettra toutefois au lecteur de se faire une idée du niveau de fiabilité atteint par DEFI sur des phrases simples ayant fait l'objet d'un contrôle lexicographique (il s'agit en effet d'exemples extraits de COBUILD).

Un dernier point. Il ne faut pas prendre DEFI pour ce qu'il n'est pas, à savoir un outil qui permettrait de repérer automatiquement des lexies en contexte. Je ne pense pas qu'un tel outil puisse être développé à l'heure actuelle. Quel que soit le soin que l'on apporte à décrire les contextes qui supportent la lexie et les manipulations dont elle peut faire l'objet, on ne pourra pas faire l'économie d'un recours au sens, au vouloir dire, lequel ne se laisse pas réduire à des corrélats linguistiques observables, en dépit de notre intuition persistante que ce devrait être le cas puisque après tout le lecteur humain ne dispose lui aussi que du texte… Thierry Fontenelle (communication personnelle) a attiré mon attention sur le groupe *free+hand* et les conditions textuelles qui permettent de distinguer la lecture phraséologique (*carte blanche*) de la lecture en tant que syntagme libre (main libre, c'est-à-dire non occupée). On aura affaire à la lexie si:

a) le groupe est au singulier
b) free est utilisé comme épithète
c) le groupe est précédé de l'article indéfini, ou, plus rarement, défini (*the free hand he had been asking for*)

Si ces conditions ne sont pas toutes réunies (pluriel *hands*, *free* en position d'attribut, groupe précédé d'un adjectif possessif) on a affaire au syntagme libre.

Il s'agit ici de conditions textuellement observables. Les exemples donnés par le BNC confirment les hypothèses dans la plupart des cas. Seul le sens permet de reconnaître le syntagme libre, et non la lexie, dans l'exemple ci-dessous (quels sont les corrélats textuels de ce sens? à supposer qu'ils existent, comment les débusquer? et si on les a débusqués pour cet exemple-ci, et puis pour bien d'autres encore, comment pourra-t-on être sûr qu'on les détient tous, puisque tout corpus est fini?):

> *Outside in the yard he reorganized his load, so he had a free hand for the torch.*

La position du lexicographe est plus confortable. Il sait que l'utilisateur ne consultera le dictionnaire que s'il a l'impression que le sens du syntagme n'est pas compositionnel. De plus, les mots *hand* et *free* sont très fréquents, et seront donc connus de beaucoup. Il n'y a dès lors pas de raison majeure d'exemplifier le syntagme *free+hand* en dehors de sa lecture phraséologique, et il n'est pas nécessaire de cerner les conditions textuelles qui permettraient de rejeter la lecture phraséologique. Il suffit de l'exemplifier, c'est-à-dire de la montrer dans ses contextes les plus fréquents. C'est ce que font nos deux dictionnaires, OH et RC.

DEFI, par son système de pondération, privilégiera les manifestations textuelles qui sont les plus proches des lexies telles que présentées par les deux dictionnaires, à savoir:

> *to have a free hand*
> *to have a free hand to do sth*
> *to give sb a free hand*

Mais il est tout prêt à 'sauter' (*silent move*) le couple *to+have* en tête des deux premières lexies, et il admettra que le déterminant *a* soit apparié à un autre élément du même type. Il admettra également que le nombre des groupes nominaux ne corresponde pas entre texte et lexie. Par là il laisse pénétrer du bruit, mais il assure également une meilleure couverture (*The announcement appeared a clear signal that Moscow was not prepared to extend to its 15 republics the free hand on domestic issues enjoyed by its Eastern European allies* – BNC s'apparie à *to have a free hand*, avoir carte blanche). Cette attitude n'a de sens que si l'utilisateur est un utilisateur humain, qui ne fait appel à l'appariement en mode *-m* que si la lecture compositionnelle lui paraît bizarre ou insuffisante.

Bibliographie

Corpus

BNC = British National Corpus (http://info.ox.ac.uk/bnc)

Dictionnaires et thésaurus

CIDE = Paul Procter (Rédacteur en chef) ([1]1995). *Cambridge International Dictionary of English*. Cambridge: CUP.

COBUILD = John Sinclair (Rédacteur en chef) ([1]1987). *Collins Cobuild English Dictionary*. London/Glasgow: Collins.

LDOCE = Paul Procter (Rédacteur en chef) ([1]1979). *The Longman Dictionary of Contemporary English*. London: Longman.

OH = Corréard, M. H. & V. Grundy (1994). *The Oxford-Hachette French Dictionary*. Oxford: OUP.

RC = Atkins, Beryl T. et al. ([4]1995). *Collins-Robert French/English English/French Dictionary*. Glasgow: HarperCollins.

WordNet = WordNet Prolog Package, téléchargeable du site Web de Princeton University. Voir aussi Miller 1990.

ROGET = ROGET'S THESAURUS, version du domaine public téléchargeable de plusieurs sites Web

Outils

Le parseur de surface ENGCG a été mis au point au département de linguistique générale de l'Université d'Helsinki. Il est commercialisé par Lingsoft Inc. (http://www.lingsoft.fi).

Prolog: Arity Prolog pour Windows: Arity Corporation, Damonmill Square, Concord, Mass.

Autres références

Bauer, D., Segond, F. & A. Zaenen (1995). "LOCOLEX: The translation rolls off your tongue." *Proceedings of the ACH-ALLC Conference*. Santa Barbara, California, 6-8.

Breidt, E., Segond, F. & G.Valetto (1996). "Local grammars for the description of multi-word lexemes and their automatic recognition in texts." F. Kiefer, G. Kiss & J. Pajzs (éd.) (1996). *Papers in Computational Lexicography - COMPLEX'96*. Linguistics Institute, Hungarian Academy of Sciences, 19-28.

Michiels, A. (1998). "The DEFI Matcher." Thierry Fontenelle et al. (éd.) (1998). *Euralex'98 Proceedings*. Liège: University of Liège, 203-11.

Michiels, A. (2000). "New Developments in the DEFI Matcher." *International Journal of Lexicography* 13(3),151-167.

Michiels, A. & N. Dufour (1998). "DEFI – a Tool for Automatic Multi-Word Unit Recognition, Meaning Assignment and Translation Selection." A. Rubio et al. (éd.) (1998). *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada. Vol. 2, 1179-1186.

Miller, G. A. (éd.) (1990). "WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3(4), 235-312.

Montemagni, S., Federici, S. & V. Pirrelli (1996). "Example-based Word Sense Disambiguation: a Paradigm-driven Approach." M. Gellerstam et al. (éd.) (1996). *Euralex'96 Proceedings*. Göteborg University, 151-160.

Segond, F. & E. Breidt (1996). "IDAREX: Description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis." A. Clas, P. Thoiron & H. Béjoint (éd.) (1996). *Lexicomatique et Dictionnairique (Actes du Colloque de Lyon – 1995)*. Montréal et Beyrouth: Aupelf-Uref et FMA, 93-104.

Weinreich, U. (1980). "Problems in the Analysis of Idioms." W. Labov & B. Weinreich (1980). *On Semantics*. University of Pennsylvania Press.