DEFI, un outil d'aide à la compréhension

Archibald Michiels

Département de langue et linguistique anglaises – Université de Liège,
3, Place Cockerill, B-4000 Liège, Belgique,
amichiels@ulg.ac.be

Abstract

DEFI acts as a filter on a bilingual dictionary (a merge of the Oxford/Hachette (OH) and Robert/Collins (RC) English-French and French-English bilinguals) to provide the user with the most likely translation(s) of the item he has requested help about.

The tasks involved are the following:

- recognition of general language multi-word units (mwu's) stored in the bilingual dictionaries. This task also includes the presentation to the user of relevant dictionary examples, because the concept of mwu is extended here to cover examples as selected and/or edited by lexicographers

- for both multi-word units and single-word lexical items, restriction of the range of translations, such restriction to be based on properties of the source text, i.e. the textual environment of the item the user has asked to get the translation of. In the best of cases, the translation that ranks highest according to the DEFI matcher is the one that is most appropriate to the context.

Mots clés/Keywords

dictionnairique / aide à la compréhension / lecture active / sélection d'acceptions

machine-readable dictionaries / reading aids / translation aids / word sense assignment

1 Introduction

DEFI est un outil d'aide à la compréhension de textes anglais destiné aux lecteurs francophones. Il établit un filtre sur le dictionnaire bilingue anglais-français pour ne retenir que les acceptions pertinentes au contexte et en donner les équivalents français, en plaçant les plus pertinents en tête.

Bien que DEFI fournisse des traductions, il ne s'agit pas d'une aide à la traduction, dans la mesure où cette dernière est un processus extrêmement complexe qui demande une

appréhension plus globale du contexte, et des possibilités de reformulation qui dépassent le cadre phraséologique du dictionnaire bilingue.

La place manque pour présenter un exposé détaillé et approfondi des mécanismes d'appariement texte-dictionnaire mis en œuvre par DEFI. Pour pallier ce manque on se permettra de référer aux documents accessibles au départ de la page consacrée à DEFI sur le Web, à savoir http:// engdep1.philo.ulg.ac.be/ michiels/ efdefi.htm.

2 Principes directeurs du développement de DEFI

DEFI est un prototype. On s'est concentré sur les problèmes fondamentaux que présente l'appariement texte-dictionnaire pour le génie linguistique, et on a négligé le développement de l'interface utilisateur. Celle-ci devrait s'intégrer dans un traitement de texte et/ou un navigateur Web. Lorsque l'utilisateur clique sur l'item qu'il ne comprend pas, l'interface transmet au programme d'appariement l'item cliqué et l'unité textuelle dans laquelle il s'insère (phrase ou partie de phrase ; on peut imaginer d'utiliser la ponctuation lourde (. ;:?! ()) pour fixer les frontières de l'unité textuelle en question). Pour l'heure, une telle interface n'existe pas. DEFI travaille sur base de fichiers qui comportent un ensemble de lignes, chaque ligne se composant de l'item cliqué et de son contexte. On trouvera un exemple de fichier d'entrée à http:// engdep1.philo.ulg.ac.be/ michiels/ textfile.htm.

Par contre, on n'a nullement négligé la question du coût computationnel de l'appariement texte-dictionnaire. Les accès aux dictionnaires et bases de données ont été optimisés et l'algorithme d'appariement a été affiné pour offrir un compromis raisonnable entre coût computationnel et qualité de l'appariement. Le banc d'essai de base de DEFI, un ensemble de 1000 phrases extraites de la base de données d'exemples de COBUILD, fait apparaître un temps de traitement de deux secondes et demie par phrase (la phase de pré-traitement, incluant l'analyse par le parseur *engcg* de Lingsoft, prend 15 millièmes de seconde par phrase). Ces résultats sont obtenus sur un PC travaillant à 733 Mhz sous Windows 98.

2.1 Hypothèse fondamentale

Il n'y a pas lieu de se prononcer ici sur la réalité linguistique des acceptions. Elles sont le résultat de la pratique lexicographique monolingue, comme les traductions sont le résultat de la pratique lexicographique bilingue. On sait qu'il n'y a pas nécessairement de parallélisme entre acceptions et traductions, puisque la découpe sémantique effectuée par les deux langues est parfois pratiquement identique (*cell*/*cellule*) mais parfois largement divergente (*dent*/*bosse#entaille*). La base pour l'établissement d'une nouvelle paire **item-traduction** est elle toute pragmatique : en tant que lexicographe, si je dis que *x* se traduit par *y*, et qu'on me présente un *x* en contexte dont la traduction la plus naturelle est *z*, je créerai une nouvelle paire *x*-*z*, et je tenterai de spécifier en quoi les contextes où *x* se traduit par *z* diffèrent des contextes où *x* se traduit par *y*. Pour ce faire, le lexicographe dispose d'un ensemble de champs destinés à recevoir de l'information métalinguistique : collocats correspondant aux diverses positions syntaxiques ouvertes par l'item, spécification d'un domaine du discours, de l'environnement syntaxique, etc.

Les seules tâches de DEFI sont les suivantes :

- associer l'item cliqué avec une unité phraséologique répertoriée dans le dictionnaire, s'il échet ;

- élaguer l'arbre des traductions et présenter les traductions retenues dans un ordre de pertinence décroissante.

Pour DEFI, il est toutefois important de noter que le dictionnaire monolingue donne des acceptions qui sont des pôles d'attraction pour le processus interprétatif de l'item en contexte, et non des interprétations toutes faites et parfaitement discrètes qu'il suffirait de déplacer du dictionnaire vers le texte par une opération intellectuelle équivalente au *copier/coller*. De même, le dictionnaire bilingue donne des traductions qu'il convient également d'adapter au contexte. En conséquence, lorsqu'on consulte le dictionnaire, on n'est pas toujours le mieux servi si on n'obtient que l'acception qui semble la plus appropriée ou la meilleure traduction en contexte. Les autres acceptions/traductions peuvent également contribuer à l'opération d'interprétation de l'item en contexte ; elles ne peuvent être rejetées que si leur distance par rapport au contexte utilisateur est nettement plus grande. Il y a là un seuil à fixer heuristiquement.

Dans le cas de la reconnaissance d'une unité phraséologique, la décision est souvent plus facile à prendre, l'élément le plus approprié se détachant plus nettement du lot. On ne s'en étonnera pas si on considère que le dictionnaire, qui décrit les conditions d'appariement d'un lexème simple ou d'une lexie complexe à une acception ou une traduction donnée, ne peut le faire que par le biais d'informations que l'utilisateur doit être capable d'observer ou de déduire du contexte dans lequel l'item apparaît hors dictionnaire. Dans le cas du lexème réduit à un seul mot, il s'agit de propriétés morphosyntaxiques de l'item (possibilité de pluriel, de formes conjuguées, de degrés de comparaison) et de spécifications de son environnement, spécifications dont le degré de précision ne va toutefois pas jusqu'au mot, mais s'arrête à la classe thésaurique (collocats) ou reste au niveau de la construction morphosyntaxique. L'unité phraséologique, quant à elle, même si elle doit tenir compte de la variabilité, est en fait un ensemble de spécifications au niveau du mot, et permet donc, pour chacun de ses constituants de base, un ancrage plus aisé à repérer hors dictionnaire.

Pour autant que l'item cliqué ou le lemme qui lui correspond figure au dictionnaire, soit comme item indépendant soit comme point d'ancrage d'une unité phraséologique, DEFI produira une ou plusieurs traductions. Toutes les propriétés calculables associées au couple item-traduction par le dictionnaire lui serviront à mesurer la pertinence de ce couple pour le contexte utilisateur. Chacune donnera lieu au calcul d'un poids qui reflète le degré de qualité avec lequel la propriété peut être associée au contexte utilisateur. Il est à noter que le non respect d'une spécification quelconque, même d'une spécification aussi essentielle que la partie du discours, ne conduira pas à l'échec de la prise en compte du couple item-traduction. Cette attitude prudente est dictée par le fait que les propriétés, y compris celles qui paraissent le mieux calculables et semblent le plus clairement relever d'un choix binaire (la propriété est ou n'est pas présente), ne sont pas calculables avec un degré de certitude suffisant pour qu'elles puissent faire barrage. La partie du discours est souvent calculée par le parseur de surface utilisé par DEFI, à savoir *engcg* de Lingsoft, de manière erronée ou insuffisante (pas de décision univoque, comme dans le cas des formes en **ing** et **ed**, qui sont attribuables à la fois au verbe et au nom, ou au verbe et à l'adjectif).

3 Le traitement de la phraséologie

DEFI tâche de prendre en compte le rôle capital joué par la phraséologie, c'est-à-dire par tous les éléments plus ou moins figés dont la taille en mots est supérieure à un. Cette définition peu orthodoxe est volontairement très large : DEFI applique le même traitement aux unités phraséologiques reconnues comme telles par la lexicographie monolingue et bilingue (*phrasal verbs*, *idioms*, etc.) et aux exemples donnés par le dictionnaire pour illustrer les acceptions qu'il répertorie. En effet, si on peut établir avec plus ou moins de justification linguistique la frontière entre les expressions semi-figées et les exemples qui tentent de replacer l'item dans son contexte le plus typique, il est à noter que le dictionnaire bilingue ne le fait pas. Une expression sera présentée dans le contexte plus large de la phrase si la traduction en est ainsi facilitée – le lexicographe ne vise pas à donner les squelettes les plus dépouillés, qui seraient souvent difficiles traduire sinon maladroitement. *I'll miss you* (*tu me manques*) est beaucoup plus facile à traduire que *to miss someone* (*regretter l'absence de quelqu'un*).

DEFI pousse le parallélisme plus loin. Le traitement qu'il applique à toute la phraséologie du dictionnaire, il l'applique également à la phrase qui contient l'item qui pose un problème de compréhension à l'utilisateur. De cette façon, il est à même de mesurer avec un plus grand degré de précision la distance qui sépare le texte utilisateur de l'unité phraséologique candidate à l'appariement.

De quel traitement s'agit-il ? DEFI soumet l'unité textuelle (phrase utilisateur, lexie complexe, exemple illustratif) au parseur de surface *engcg* développé par Lingsoft ; ensuite le programme *tagtxt* (une application *awk*) construit sur les résultats du parseur pour tenter d'approfondir l'analyse très en surface que fournit *engcg*. Il tente notamment d'établir une liste de groupes nominaux (en spécifiant également les constituants têtes de ces groupes) qui permettront un appariement phraséologie-texte dans le cas de *fillers* lexicographiques tels que *something* ou *somebody*. Il débusque aussi les relations syntaxiques auxquelles peuvent participer les collocats : sujet, objet direct, complément du nom, etc. pour rendre possible une mesure de la distance qui sépare le collocat prévu dans l'unité lexicographique et l'élément qui remplit le rôle syntaxique voulu dans le texte de l'utilisateur. Ces relations syntaxiques doivent tenir compte des 'transformations' de l'ordre canonique des éléments telles que la passivisation. Finalement, DEFI émet une hypothèse structurelle sur l'unité phraséologique tout entière : s'agit-il d'une phrase, d'un groupe verbal, nominal, prépositionnel ? Il faut en outre calculer la polarité (affirmatif v. non-affirmatif) pour pouvoir traiter adéquatement les unités phraséologiques qui ont une négation inhérente (*not give a damn* / *ne pas faire dans la dentelle*). En effet, cette négation peut prendre une forme différente de celle spécifiée dans l'unité phraséologique, ou encore avoir migré vers une proposition supérieure dans la hiérarchie syntaxique (*I doubt whether he would give a damn* / *il n'a jamais donné l'impression de faire dans la dentelle*).

Tagtxt a aussi pour mission de donner un poids à tous les traits morphosyntaxiques attribués par le parseur aux constituants du texte. Le programme d'appariement texte-dictionnaire aura soin de collecter les traits qui se correspondent dans le texte de l'utilisateur et l'unité phraséologique candidate à l'appariement. L'accumulation du poids de ces traits jouera un rôle dans la mesure de la qualité de l'appariement.

On trouvera un exemple d'analyse de *engcg*, et son enrichissement par *tagtxt*, dans http:// engdep1.philo.ulg.ac.be/ michiels/ parse.htm.

L'approche préconisée par DEFI s'oppose assez nettement à la spécification de grammaires locales qui détermineraient avec précision le degré et la nature de la variabilité acceptable pour une unité phraséologique donnée, et les divers points d'insertion pour des éléments étrangers à l'unité elle-même. Cette deuxième approche caractérise le projet **Locolex** de Rank Xerox ((Bauer et al., 1995), (Breidt et al., 1996), (Segond et Breidt, 1996)). On se référera à http:// engdep1.philo.ulg.ac.be/ michiels/ frmwu.htm. L'argument de base est que, en anglais surtout, les expressions figées sont fréquemment dégelées dans un processus allusif ; la condition suffisante est que l'expression de base soit perceptible sous les modifications. Il ne semble pas possible de rendre compte de ces modifications par le biais de grammaires locales, qui devraient alors être trop 'relâchées' pour être encore utiles. DEFI se contente de mesurer la distance qui sépare l'expression répertoriée dans le dictionnaire de son avatar textuel, sans poser de conditions du type tout ou rien.

4 Ressources lexicales

En ce qui concerne les ressources lexicales mises en oeuvre, DEFI a fusionné en un seul dictionnaire bilingue ses deux bilingues source, à savoir RC et OH, qui sont à mes yeux les deux fleurons de la lexicographie bilingue anglais-français français-anglais. Seule la direction anglais→ français est exploitée. La fusion s'est faite d'une manière très conservatrice, sans éviter toute redondance mais en assurant qu'aucune information ne se perde (se référer à http:// engdep1.philo.ulg.ac.be/ michiels/ merge.htm). Elle a considérablement accru le pouvoir discriminateur de certains champs, en particulier le champ des collocats. Les entrées ont été réorganisées en couples lemme-traduction, avec répartition de l'information métalinguistique dans différents champs. Ce dictionnaire a deux formes directement consultables par le développeur DEFI: sous sa forme *defidic* (cf. http:// engdep1.philo.ulg.ac.be/ michiels/ rissh33.htm), il se présente comme une série d'enregistrements *awk*, directement accessibles via *awk* ou un éditeur. Sous la forme de termes *Prolog*, il est consultable au moyen du logiciel *lkp*, une application *Prolog* développée dans le cadre du projet DEFI (cf. http:// engdep1.philo.ulg.ac.be/ michiels/ lkpextr.htm).

Le programme d'appariement texte-dictionnaire, qui est au centre de DEFI, utilise lui deux dictionnaires résultant d'une transformation de *defidic*. Le dictionnaire *dic* contient toutes les unités phraséologiques du dictionnaire de départ (au sens très large exposé ci-dessus) ; *sdic* est le dictionnaire des unités lexicographiques qui ne dépassent pas la frontière du mot. Toutes les unités phraséologiques ont été soumises au parseur *engcg* et ensuite à *tagdic*, une application *awk* tout à fait similaire et parallèle à *tagtxt*. On trouvera un extrait de *dic* à http:// engdep1.philo.ulg.ac.be/ michiels/ dic.htm, et de *sdic* à http:// engdep1.philo.ulg.ac.be/ michiels/ sdic.htm.

DEFI utilise trois dictionnaires monolingues de l'anglais, tous trois s'inscrivant dans une perspective pédagogique, ce qui garantit le caractère relativement explicite des informations qu'ils fournissent pour distinguer les diverses acceptions, et la simplicité de leur vocabulaire définitoire ainsi que la typicité de leurs exemples, qu'il s'agisse d'exemples forgés (LDOCE) ou extraits d'un corpus (COBUILD, CIDE) et présentés tels quels ou retravaillés pour augmenter leur généralité. L'appel aux monolingues se fait dans le cadre du traitement du champ *Indicateur*, un champ du dictionnaire bilingue où le lexicographe donne au lecteur des pointeurs qui lui permettent de reconnaître la pertinence de la traduction proposée pour le contexte utilisateur, pointeurs qu'il ne peut formaliser dans le cadre des autres champs qui servent également à cerner les traits pertinents du contexte, comme par exemple les collocats ou les étiquettes matière. On trouvera des extraits des dictionnaires monolingues sous la forme

de termes *Prolog* directement utilisables par DEFI dans http:// engdep1.philo.ulg.ac.be/ michiels/ mono.htm.

De plus, DEFI fait appel à deux bases de données qu'on peut qualifier de thésauriques, *Roget* et *Wordnet*. *Roget* est plus nettement littéraire, et sa base est souvent la simple association d'idées plutôt que de strictes relations thésauriques telles que l'hyponymie, la synonymie et l'antonymie, relations de base pour *Wordnet*.

Enfin, DEFI se sert d'une base de données extraite du dictionnaire bilingue fusionné. Cette base de données rend compte des relations de cooccurrence entre collocats, et est appelée dans le calcul de la distance qui sépare l'item apparaissant dans le texte utilisateur du collocat spécifié par le dictionnaire (hypothèse de Montemagni et al. 1996).

5 Utilisation des champs du dictionnaire bilingue

Tous les champs du dictionnaire contribuent à déterminer si un couple item-traduction est pertinent dans le contexte utilisateur. Mais au départ ils sont tous destinés au lecteur humain et non à l'ordinateur. Ce n'est pas parce qu'un champ est formalisé qu'il est utilisable systématiquement et sans risque d'erreur par un programme d'appariement tel que celui qui est au cœur de DEFI. Il suffit de penser au champ **étiquette matière** (*field label*). Il n'est calculable avec fiabilité que sur un contexte beaucoup plus large que la phrase (l'unité de travail de DEFI, pour des raisons évidentes de coût computationnel). De plus, il est difficile d'éviter les pièges tendus par les ruptures d'isotopie et les métaphores, qui glissent d'un domaine à un autre.

De plus, l'utilisation de tous les champs d'information présuppose que l'on a affaire à un texte **interprété**, pas seulement **analysé** en structures de surface. Il est indéniable que le parseur rend un grand service en transformant une liste de mots en un ensemble de couples variante morphologique – forme lemmatisée, et qu'il procure à *tagtxt* et *tagdic* des éléments d'analyse sur lesquels ces derniers peuvent envisager de construire les relations syntaxiques dont le traitement des collocats a besoin. Mais le risque d'erreur est présent partout.

En conséquence, le poids que DEFI attribuera à chaque champ dépendra non seulement du pouvoir discriminateur de ce champ, mais aussi du degré de fiabilité avec lequel la propriété décrite par le champ peut être mesurée. Par exemple, les collocats (sujets et objets typiques) ont un très grand pouvoir discriminateur (l'utilisateur humain en fait grand cas) mais leur nature en rend le traitement assez délicat. Il ne s'agit pas de variantes morphosyntaxiques ou encore de lemmes, mais bien de têtes thésauriques, qu'on ne pourra exploiter qu'en faisant appel à une organisation thésaurique du lexique (*Roget*, *Wordnet*) et aux liens tissés par la cooccurrence à l'intérieur même du domaine.

Quant au champ **Indicateur**, le fourre-tout pour toutes les propriétés discriminatoires qui ne trouvent pas place ailleurs dans l'entrée de dictionnaire, il est hautement intéressant, mais le traitement, de par la diversité des indications et le caractère peu structuré de leur présentation, est extrêmement difficile. On se référera à (Michiels 2000).

DEFI permet au développeur de spécifier les poids attribués aux différentes propriétés dans son programme d'appariement. Cette souplesse est bienvenue dans la perspective de la mise au point du programme, mais il faut se garder d'apporter des modifications aux heuristiques de poids sur base de l'une ou l'autre phrase pour lesquelles DEFI performe médiocrement. A

chaque modification, il convient de réappliquer les programmes à une série de phrases tests qui reprennent un assez grand nombre de cas d'appariement. Il faut aussi se garder de tenter de 'corriger' des erreurs du parseur par des modifications de ce type. DEFI est dépendant d'un traitement en amont – il faut s'y résigner.

Dans une perspective à plus long terme, on peut envisager d'introduire des techniques d'apprentissage automatique. On peut indiquer à DEFI quelles sont les traductions à privilégier, et lui laisser le soin de découvrir la pondération qui y conduit. Ici encore, un tel apprentissage n'a de chance de succès que s'il prend pour base un univers suffisant, vraisemblablement plusieurs dizaines de milliers de phrases tests. Se pose alors le problème de la détermination des traductions à privilégier – pour servir de benchmark, celle-ci doit bien sûr être indépendante des résultats fournis par DEFI.

6 Evaluation

L'évaluation d'un outil tel que DEFI est très ardue. Tout d'abord parce que le domaine d'application est infini : DEFI se propose de donner une traduction pour tout item (à l'exception des mots outils les plus fréquents) appartenant à un texte rédigé en anglais. Quel que soit l'échantillon d'évaluation, il sera toujours infime par rapport à l'univers qu'il veut refléter. En second lieu, la pertinence des choix proposés par DEFI ne peut être mesurée que par un utilisateur humain, ce qui réduit à nouveau la taille des échantillons que l'on peut raisonnablement traiter et introduit un important facteur de subjectivité. Il est en effet difficile de se mettre d'accord sur un classement des traductions proposées, même s'il est assez aisé de débusquer les erreurs grossières (mais le sont-elles toujours vraiment ? une erreur sur la partie du discours peut s'avérer moins grave qu'un léger glissement dans l'évaluation des collocats, pour autant que le sémantisme de base soit préservé et l'interprétation aisément dérivable d'une partie de discours vers une autre).

On peut envisager trois types de fichiers tests pour DEFI:

1) Une suite de phrases tests utilisée par le développeur. Elle peut comprendre des phrases inventées destinées à tester telle ou telle fonction du programme d'appariement. On aura soin de soumettre à nouveau cet ensemble à chaque modification apportée au programme, pour s'assurer que les gains engrangés dans le traitement de *x* ne soient pas reperdus dans le traitement de *y*.

2) Des tests basés sur les besoins réels d'utilisateurs réels. Un seul test de ce type a été réalisé, avec des étudiants en anglais de première et de dernière année à l'Université de Liège. On trouvera le fichier test dans http:// engdep1.philo.ulg.ac.be/ michiels/ firstres.htm.

3) Un banc d'essai 'privilégié'. Il s'agit de phrases servant d'exemples illustratifs dans les dictionnaires monolingues. Ce banc d'essai est privilégié dans la mesure où les exemples de dictionnaire sont censés présenter les contextes d'utilisation les plus typiques, ceux qui exercent le plus grand pouvoir d'attraction vers une acception donnée. DEFI devrait offrir ici de meilleures performances que sur du texte brut, et les résultats devraient conduire plus rapidement à des améliorations du programme d'appariement. Un tel banc d'essai est constitué par un millier de phrases d'exemples extraites de COBUILD – j'y ai déjà fait référence dans l'indication des performances de DEFI. On trouvera le fichier complet, avec indication de la traduction sélectionnée par un utilisateur humain (l'auteur de cette communication), dans

http:// engdep1.philo.ulg.ac.be/ michiels/ cobres.htm. L'appendice à ce document en donne un bref extrait (une dizaine de phrases contenant toutes une lexie).

Appendice

La flèche (→) indique la traduction correcte.

/* 1 /

[72] /* numéro de l'exemple dans la base de données */

Restrictive practices were putting a savage brake on enterprise. /* exemple */

Clwlist = [brake] /* item pour lequel l'aide est demandée */

Processing_time(0,0,0,44) /* temps de traitement – heures, minutes, secondes, centièmes de seconde (temps réel) */

→139 - 309307, /* pondération (139) et numéro d'identification de la lexie dans la base de données DEFI */

efm, / * origine : entrée résultant de la fusion (m=merged) d'une entrée de RC et d'une entrée d'OH */
[brake],
to put the brakes on, /* lexie sélectionnée */
tr(freiner), /* traduction de la lexie */

[m(c1,vac,to,0),
m(c6,dic(put),txt(putting),morph(0),syn(3),30),
m(c17,dic(the),txt(a),morph(5),syn(0),5),
m(c6,dic(brakes),txt(brake),morph(5),syn(5),30),
m(c5,dictxt(on),morph(0),syn(0),50)]

/* compte rendu du parcours de la lexie : *to put the brakes on – to* a fait l'objet d'un *silent move* dans le parcours de la lexie ; *putting* correspond à *put* ; *the* est apparié à *a* par partage de classe (classe des déterminants) ; *brakes* correspond à *brake* et *on* à *on* ; le résultat d'appariement de traits morphologiques/syntaxiques se trouve en argument du foncteur *morph/syn*. Le dernier argument du foncteur *m* est le poids global attribué à la transition. */

/* 2 */
[106] Stevens promptly notified the German authorities through the normal channels.
Clwlist = [channels]
Processing_time(0,0,0,50)
→249 - 287334, ohef, [channels], to do sth through the normal channels, tr(faire qch par la voie normale),
[m(c1,vac,to,0),m(c22,dic(dosth),jmpd,0),m(c5,dictxt(through),morph(2),syn(8),50),m(c5,dictxt(the),morph(9),syn(0),25),m(c5,dictxt(normal),morph(5),syn(3),50),m(c5,dictxt(channels),morph(6),syn(3),50),m(c4,vac,punct,0)]

/* 3 */
[119] Now that I've told her everything, I can leave with a clear conscience.
Clwlist = [clear]
Processing_time(0,0,3,18)
→263 - 287364, ohef, [clear], to do sth with a clear conscience, tr(faire qch la conscience tranquille),
[m(c1,vac,to,0),m(c20,dic(dosth),txt(leave),15),m(c5,dictxt(with),morph(2),syn(5),50),m(c5,dictxt(a),morph(10),syn(0),25),m(c5,dictxt(clear),morph(5),syn(3),50),m(c5,dictxt(conscience),morph(7),syn(3),50),m(c4,vac,punct,0)]
235 - 118773, rcef, [clear], he left with a clear conscience, tr(il est parti la conscience tranquille),
[m(c17,dic(he),txt(i),morph(7),syn(5),5),m(c6,dic(left),txt(leave),morph(5),syn(0),30),m(c5,dictxt(with),morph(2),syn(5),50),m(c5,dictxt(a),morph(10),syn(0),25),m(c5,dictxt(clear),morph(5),syn(3),50),m(c5,dictxt(conscience),morph(7),syn(3),50),m(c4,vac,punct,0)]

/* 4 */
[299] It fell to Philip Crow to act the part of host.
Clwlist = [fell]
Processing_time(0,0,2,86)
→137 - 144389, ohef, [fell], it falls to sb to do, tr(c'est à qn de faire, c'est à qn qu'il incombe de faire {fml}),

[m(c5,dictxt(it),morph(6),syn(5),25),m(c6,dic(falls),txt(fell),morph(7),syn(3),30),m(c5,dictxt(to),morph(2),syn(5),25),
m(c12,dic(sb),txt(crow),12),m(c5,dictxt(to),morph(2),syn(0),25),m(c15,dic(do),txt(act),12)]

/* 5 */
[499] She was dressed in a yellow sari with yellow ribbons to match in her hair.
Clwlist = [match]
Processing_time(0,0,2,20)
→163 - 345130, rcef, [match], with a skirt to match, tr(avec (une) jupe assortie),
[m(c5,dictxt(with),morph(2),syn(5),50),m(c3,vac,np(1,5,c(2,3)),0),m(c5,dictxt(to),morph(2),syn(0),25),m(c5,dictxt(match),morph(7),syn(6),50)]

/* 6 */
[524] They were all cast in the same contemporary mould, with flowing shoulder-length hair.
Clwlist = [mould]
Processing_time(0,0,1,37)
247 - 138786, ohef, [mould], in the same mould, tr(dans le même moule),
[m(c5,dictxt(in),morph(2),syn(5),50),m(c5,dictxt(the),morph(9),syn(0),25),m(c5,dictxt(same),morph(5),syn(3),50),m(c5,dictxt(mould),morph(7),syn(3),50)]
→239 - 274867, ohef, [mould], to be cast in a mould, tr(être coulé dans un moule),
[m(c1,vac,to,0),m(c6,dic(be),txt(were),morph(5),syn(0),30),m(c5,dictxt(cast),morph(2),syn(3),50),m(c5,dictxt(in),morph(2),syn(5),50),m(c17,dic(a),txt(the),morph(5),syn(0),5),m(c5,dictxt(mould),morph(7),syn(3),50)]

/* 7 */
[625] Everything was in its proper place.
Clwlist = [proper]
Processing_time(0,0,1,16)
→240 - 91197, ohef, [proper], everything is in the proper place, tr(tout est à sa place),
[m(c5,dictxt(everything),morph(6),syn(5),50),m(c6,dic(is),txt(was),morph(7),syn(3),30),m(c5,dictxt(in),morph(2),syn(5),50),m(c17,dic(the),txt(its),morph(0),syn(0),5),m(c5,dictxt(proper),morph(5),syn(3),50),m(c5,dictxt(place),morph(7),syn(3),50)]

/* 8 */
[719] He was not easily ruffled.
Clwlist = [ruffled]
Processing_time(0,0,0,28)
→144 - 222177, rcef, [ruffled], she wasn't at all ruffled, tr(elle était parfaitement calme),
[m(c17,dic(she),txt(he),morph(8),syn(5),5),m(c5,dictxt(was),morph(9),syn(3),50),m(c5,dictxt(not),morph(5),syn(0),25),m(c18,dic(at=all),txt(easily),morph(5),syn(5),5),m(c5,dictxt(ruffled),morph(2),syn(3),50)]

/* 9 */
[851] That school has always had a bit of a struggle to keep going.
Clwlist = [struggle]
Processing_time(0,0,2,91)
→160 - 266376, ohef, [struggle], they had a struggle to do, tr(ils ont eu du mal à faire),
[m(c17,dic(they),txt(that),morph(2),syn(0),5),m(c6,dic(had),txt(has),morph(7),syn(0),30),m(c5,dictxt(a),morph(10),syn(0),25),m(c5,dictxt(struggle),morph(7),syn(5),50),m(c5,dictxt(to),morph(2),syn(0),25),m(c15,dic(do),txt(keep),12)]

/* 10 */
[882] I think I would like to take Tony up on something that he said.
Clwlist = [take,up,on]
Processing_time(0,0,1,15)
→304 - 12287, rcef, [take,up,on], I would like to take you up on something you said earlier, tr(je voudrais revenir sur quelque chose que vous avez dit précédemment),
[m(c25,dic(i,would),jmpd,0),m(c5,dictxt(like),morph(7),syn(3),50),m(c5,dictxt(to),morph(2),syn(0),25),m(c5,dictxt(take),morph(7),syn(3),50),m(c3,vac,np(0,1,c(0,1)),0),m(c5,dictxt(up),morph(0),syn(5),50),m(c5,dictxt(on),morph(0),syn(5),50),m(c5,dictxt(something),morph(6),syn(3),25),m(c3,vac,np(0,1,c(0,1)),0),m(c5,dictxt(said),morph(8),syn(3),50),m(c19,dic(earlier),crossdicpos,0),m(c4,vac,punct,0)]

Références

Dictionnaires et thésaurus

CIDE = Paul Procter, Rédacteur en chef, (1995), *Cambridge International Dictionary of English*, Cambridge University Press (première édition)

COBUILD = John Sinclair, Rédacteur en chef, (1987), *Collins Cobuild English Dictionary*, Collins (première édition)

LDOCE = Paul Procter, Rédacteur en chef, (1979), *The Longman Dictionary of Contemporary English* (première édition)

OH = M.H. Corréard et V. Grundy, (1994) *Oxford-Hachette French Dictionary*, Oxford University Press

RC = Beryl T. Atkins et al. '1995) *Collins-Robert French/English English/French Dictionary*,Glasgow: HarperCollins (4ème édition)

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).

Awk: implantations pour Windows de MKS et Thompson; voir également (Aho et al., 1988)

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

Autres références

Aho, A.V., Kernighan, B.W. et Weinberger, P.J. (1988), *The AWK Programming Language*, Addison-Wesley, Reading, Mass.

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

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

Michiels, A. (2000), 'New Developments in the DEFI Matcher', *International Journal of Lexicography*, Vol 13, Nr 3.

Miller, G. A., (ed) (1990), 'WordNet: An On-Line Lexical Database', *International Journal of Lexicography*, Volume 3, Nr 4.

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

Segond, F. and Breidt, E. (1996), 'IDAREX: Description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis', in Clas, Thoiron and Béjoint (eds) *Lexicomatique et Dictionnairique (Actes du Colloque de Lyon – 19 95)*, Aupelf-Uref and FMA, Montréal and Beyrouth, pp.93-104.