CoBRA

Deville, Guy; DUMORTIER, Laurence; Meurisse, Jean-Roch

Published in:
Penser, écrire et interpréter le droit

Publication date:
2022

Document Version
le PDF de l'éditeur

CoBRA : un outil raisonné d'aide à la lecture de textes juridiques en anglais et en néerlandais

Guy Deville

École des langues vivantes-UNamur

Laurence Dumortier

et

Jean-Roch Meurisse

Cellule TICE-UNamur

Introduction

Lors de la lecture d'un texte en langue étrangère, nombreux sont ceux qui recourent à un dictionnaire bilingue en ligne pour en faciliter la compréhension. Cette ressource utile – voire indispensable – demeure toutefois non intégrée à l'activité de lecture même : le choix de la traduction appropriée d'un mot ou d'une expression en fonction de son contexte original est laissé à l'initiative du lecteur, qui doit sélectionner les informations pertinentes (catégorie syntaxique, sens, emploi).

D'expérience, nous observons qu'une telle tâche n'est pas à la portée immédiate des étudiants de l'enseignement universitaire, où la lecture de textes constitue le support privilégié de l'apprentissage d'un vocabulaire technique lié à un domaine de spécialité. Tel est le cas de la terminologie juridique anglaise et néerlandaise. À notre connaissance, il n'existe pas aujourd'hui d'applications multimédias, ni de ressources en ligne intégrées et spécifiquement dédiées à la lecture de textes en langue étrangère de niveau spécialisé à l'usage des apprenants francophones, qui proposent (i) une traduction contextualisée de tous les mots et expressions de ces textes et (ii) des phrases extraites de très grands corpus multilingues alignés (également appelés *bi-textes*) pour illustrer ces traductions sous forme de concordances bilingues.

Ces raisons nous ont amenés à mettre en œuvre un outil qui rencontre les fonctionnalités décrites ci-dessus. Intégré à la plateforme d'e-learning *WebCampus* de l'Université de Namur, l'outil CoBRA (*Corpus-Based Reading Assistant*) propose pour tout texte d'un cours de langues (anglais et néerlandais) un accès par un clic à la traduction française de chaque mot ou expression. Cette traduction est illustrée à l'aide de concordances issues de très grands corpus bilingues alignés qui recouvrent la langue usuelle ainsi que différents domaines de spécialité. Les exemples de traduction de chaque mot sont toujours fournis en fonction du contexte de ce dernier dans le texte de la leçon, les ambiguïtés (syntaxiques ou sémantiques) ayant été levées lors de la préparation du texte pour sa mise en ligne (phase appelée *balisage* ou *étiquetage*).

Le présent article reprend d'abord la genèse du projet CoBRA et ses motivations, qui sont nées de la mise en œuvre de cours de langues pour le programme de bachelier en droit à l'UNamur. Il décrit ensuite la suite d'outils intégrés à CoBRA, et en particulier le module de lecture ainsi que l'outil de balisage de textes. Une attention particulière est accordée à la structure du lexique de CoBRA, pierre angulaire dans la création et la consultation des textes.

SECTION 1. – Historique et motivation du projet

Le projet CoBRA est le fruit d'une vingtaine d'années de recherche et de développement au sein de l'École des langues vivantes (ELV) et de la Cellule des technologies de l'information et de la communication au service de l'enseignement (TICE) de l'Université de Namur[1]. Il s'appuie sur une longue pratique de l'enseignement du néerlandais et de l'anglais à des apprenants francophones non spécialistes (qui ne suivent pas une filière *langues et littératures étrangères*) en milieu universitaire.

D'expérience, on observe que l'opacité du vocabulaire d'une langue étrangère, sa structure morpho-lexicale très éloignée du français (c'est le cas du néerlandais) ou sa dimension très idiomatique (c'est le cas de l'anglais) sont les principales pierres d'achoppement dans la maîtrise de ces langues pour tout apprenant francophone.

Le lexique revêt dès lors une dimension importante dans les méthodes d'apprentissage de ces langues destinées à cette catégorie d'apprenants.

[1] G. Deville, L. Dumortier et H. Paulussen, *Génération de corpus multilingues dans la mise en œuvre d'un outil en ligne d'aide à la lecture de textes en langue étrangère*, 7es Journées internationales d'analyse statistique des données textuelles – JADT, Louvain-la-Neuve, 2004.

Les manuels ou syllabus d'apprentissage des langues étrangères à l'usage des étudiants de l'enseignement universitaire incluent généralement une série de textes accompagnés de leur vocabulaire. Ce vocabulaire résulte d'un choix de l'auteur, et est souvent présenté en listes bilingues de mots isolés ou repris dans un contexte minimal. Les limites d'une telle présentation statique sont évidentes : (i) pour déchiffrer un texte, l'apprenant doit constamment passer physiquement du texte à la liste de vocabulaire et *vice versa*, (ii) le vocabulaire proposé ne couvre qu'un sous-ensemble des mots du texte, qui a été arrêté unilatéralement, et (iii) une fois créée, cette liste unique est figée, car son contenu ne peut varier selon le niveau de connaissance de l'apprenant.

Les versions en ligne des cours de langues reprennent généralement un dictionnaire bilingue, consultable à la demande de l'étudiant. Cette approche supprime les inconvénients liés au support écrit : elle offre l'accès à un dictionnaire avec une couverture lexicale maximale des textes, et la fréquence de consultation de ce dictionnaire est fonction du niveau de l'étudiant. Toutefois, le dictionnaire électronique est un outil non intégré : il est toujours consulté hors contexte, car le choix de la traduction d'un mot est laissé à l'initiative de l'apprenant, qui doit sélectionner dans le dictionnaire les informations appropriées (catégorie grammaticale, sens, exemples) correspondant au contexte original. Associer un mot inconnu rencontré lors de la lecture d'un texte à une entrée de dictionnaire correspondante est une démarche malaisée qui n'est pas à la portée immédiate de tous les apprenants.

Les concordanciers multilingues – le plus souvent bilingues – en ligne, tels *TransSearch*[2] ou *Linguee*[3], proposent à l'utilisateur, sur la base d'une requête (mot ou expression), une liste de concordances bilingues (phrases ou paragraphes) extraites de très grands corpus multilingues qui ont été préalablement alignés (où chaque phrase ou paragraphe a été mis en correspondance avec sa traduction). Les concordances proposées illustrent le mot ou l'expression dans toutes les acceptions possibles, en retournant une liste qui peut être triée selon les différentes traductions de ce mot. Ici aussi, nous pouvons faire la même observation qu'avec les dictionnaires destinés à la traduction : les concordances bilingues proposées par de tels outils ne tiennent pas compte du contexte d'utilisation original du mot-clé.

---

[2] J. Bourdaillet, S. Huet, Ph. Langlais et G. Lapalme, « TransSearch: from a Bilingual Concordancer to a Translation Finder », Machine Translation, 2010, vol. 24, nos 3-4, pp. 241-271.

[3] Linguee, http://www.linguee.com (21 septembre 2012). Longman, *Longman Dictionary of Contemporary English*, Longman, 2011.

Ces constats ont amené les auteurs à concevoir CoBRA, qui est un outil original de création de supports en ligne d'aide à la lecture de textes en néerlandais et en anglais. Cet outil offre une aide raisonnée aux apprenants en les guidant dans leur recherche de la traduction exacte des mots inconnus, en illustrant ceux-ci dans le sens recherché. L'apprenant peut également constituer et exporter un glossaire personnel qui est généré à partir d'un ou de plusieurs texte(s) de leçon.

SECTION 2. – CoBRA : un outil à deux versants

CoBRA est conçu comme un outil générique interactif à deux versants, dont le premier (à l'usage des enseignants) est dédié à la création de ressources : il permet (i) de rassembler un ensemble de textes en anglais et en néerlandais (une *collection* dans notre terminologie), (ii) d'étiqueter (*baliser* dans notre terminologie) chacun des mots ou expressions de ces textes, (iii) d'ajouter au dictionnaire les entrées ou emplois manquants, (iv) de gérer des corpus bilingues préalablement alignés au niveau de la phrase et (v) de générer automatiquement des concordances extraites de ces corpus pour illustrer les entrées et emplois.

Le second versant met ces ressources à la disposition des apprenants. Il offre une série de textes ainsi préalablement balisés, où l'apprenant peut interroger chaque mot ou expression d'un clic pour en connaître la traduction dans le contexte de lecture. Cette traduction fait l'objet d'une série d'illustrations sous la forme de citations bilingues. Ces textes constituent un type de ressource dans *WebCampus*, la plateforme *Moodle* LMS de l'UNamur. Une fonction permet à l'apprenant de se constituer un *glossaire* à la carte. Enfin, un dictionnaire bilingue *néerlandais-français* et *anglais-français* est également proposé.

Nous allons détailler ces deux versants de l'outil CoBRA respectivement spécialisés dans l'utilisation et la création de ressources lexicales, après avoir décrit en bref la structure et la motivation du lexique CoBRA qui assure la cohésion entre ces deux versants.

§ 1. Lexique CoBRA

Le lexique CoBRA est une base de données dans laquelle une série de tables permettent de faire le lien entre les lexiques français, néerlandais et anglais, et d'obtenir deux dictionnaires bilingues distincts (*néerlandais-français* et *anglais-français*). À quelques particularités près liées à la langue

source, ces deux dictionnaires – qui sont en réalité des vues bilingues de cette base de données lexicale – présentent une structure comparable que nous décrivons dans cette section.

Le lexique CoBRA est constitué d'entrées simples (ou *lemmes*) et d'entrées complexes[4] (ou *locutions*). Un lemme est défini comme un item constitué d'une et une seule chaîne de caractères, dont la catégorie principale peut être *adjectif, adverbe, conjonction, déterminant, interjection, nom, verbe, préposition*. Cette définition opératoire englobe donc pour l'anglais des mots tels que (adv) *annually* (annuellement, tous les ans) ou (n) *apneoa* (apnée), mais aussi (adj) *car-free* (sans voiture) ou (n) *cycle-route* (piste cyclable).

Les entrées complexes sont constituées d'au moins deux lemmes séparés par un espace, à l'exception des locutions latines (*de facto*) et entités nommées (*Los Angeles*), qui seront reprises en tant que lemme. Ces entrées complexes reprennent les locutions de toutes les catégories syntaxiques précitées (*adverbiale*, *nominale*, *verbale*…) et les verbes complexes (*phrasal verbs* anglais et verbes séparables néerlandais).

Dans le cas de locutions en anglais, nous avons adopté une approche pragmatique dans la définition des conditions qui autorisent leur insertion dans le lexique. Sont reprises en principe dans le lexique CoBRA seules les locutions qui répondent à un des critères suivants :

(i) la locution revêt un caractère idiomatique, son sens est non compositionnel, c'est-à-dire qu'il ne peut être directement dérivé de la somme du sens de chacun de ses composants ;

(ii) la traduction française privilégiée d'une locution ne peut être uniquement dérivée de la traduction de chacun de ses composants.

Ainsi, les entrées telles que *class action* (recours collectif), *heart attack* (infarctus, crise cardiaque) ou *internet user* (internaute), *(to) take part to* (participer à) seront reprises au lexique CoBRA, tandis que les expressions suivantes n'en feront pas partie : *metal detector* (détecteur de métaux), *maple syrup* (sirop d'érable) ou *orange juice* (jus d'orange). On l'aura compris, notre démarche est inspirée par l'usage du lexique CoBRA, qui ne reprend l'emploi (et donc la traduction) de toute locution – quelle que soit sa catégorie syntaxique – que si cette traduction constitue une valeur ajoutée aux yeux des apprenants francophones.

À toute entrée de type *lemme* ou *entrée complexe* du lexique CoBRA, sont associées des informations morpho-syntaxiques liées à chaque catégorie telles que formes fléchies, genre et nombre (substantifs), caractère régulier ou irrégulier (verbes). S'il s'agit d'un lemme, on y indique un lien vers

[4] Dans le texte, les termes *mots* et *expressions* réfèrent respectivement à *lemmes* et *entrées complexes*.

ses éventuelles variantes orthographiques, et vers les entrées complexes utilisant ce lemme. À l'exception du néerlandais, ces informations ont été produites à partir de données normalisées, à savoir CELEX pour l'anglais[5] et TLF pour le français[6].

Chaque lemme ou entrée complexe faisant ainsi l'objet d'une entrée lexicale est distingué selon ses différents sens (*emplois* dans notre terminologie) s'il est polysémique. Un emploi donne donc lieu à une seule signification sous la forme (i) d'une traduction, éventuellement accompagnée de synonymes, ainsi qu'au besoin (ii) une définition ou un commentaire rédigés dans la langue source ou en français. Cette dernière option se révèle utile dans le cas d'entrées difficilement traduisibles telles que, par exemple :

– *should* (verbe modal) – *Commentaire* : verbe modal qui exprime (i) une suggestion (*you should read that book*), (ii) un conseil, une recommandation morale (*you should help her*), une opinion personnelle ou une déduction logique, (iii) une supposition (*he should be at home*) ;
– *name and shame* (locution verbale) – *Commentaire* : publication de l'identité d'une personne coupable d'un crime ou d'un comportement antisocial dans le but de provoquer chez cette personne un sentiment de honte et de susciter des remords.

Cette spécificité de notre lexique démarque celui-ci de la plupart des dictionnaires existants qui ne distinguent pas formellement tous les emplois et les traductions de chaque entrée, et où les entrées complexes sont reprises aux entrées des lemmes principaux qui les constituent.

La création d'une entrée lexicale dans CoBRA répond essentiellement à une motivation didactique : la distinction entre différents emplois pour une entrée polysémique, telle que l'entrée *duty* (Fig. 2, *infra*), s'inspire de plusieurs dictionnaires bilingues que nous confrontons, et dont nous avons omis les emplois les plus rares, tels que les acceptions obsolètes ou à caractère hautement technique (sauf pour les besoins éventuels d'un cours de nature explicitement terminologique). Seuls les principaux synonymes d'un emploi sont retenus, et les acceptions extrêmes (très formelles ou très informelles, voire argotiques) sont écartées. De même, on ne distinguera pas systématiquement le sens premier d'un terme de son sens figuré (*métonymie*) s'il n'y a pas de glissement sémantique vers une nouvelle acception qu'il conviendrait d'isoler pour des raisons didactiques.

[5] H. Baayen, R. Piepenbrock et L. Gulikers, *The CELEX Lexical Database (Release 2)*, CD-ROM, Linguistic Data Consortium, Pennsylvania (USA), University of Pennsylvania, 1995.

[6] B. Quemada et J.M. Pierrel, *Trésor de la langue française informatisé* (CD-ROM), Paris, ATILF-CNRS, 2005.

Ainsi l'entrée *diamond* (diamant) renvoie tant au sens (*emploi* dans notre terminologie) [matière] qu'à celui de [bijou]. C'est une approche contrastive avec le français qui a guidé les concepteurs.

En focalisant l'attention de l'apprenant sur l'emploi d'un terme dans son contexte de lecture par le véhicule de ses principales traductions qui sont contextualisées au moyen de concordances bilingues, nous voulons éviter de le noyer dans une masse d'informations (sens, traductions, exemples) qui répondrait au seul souci (légitime) d'exhaustivité d'une démarche strictement lexicographique.

On notera que le versant néerlandais du lexique décrit ci-dessus a été essentiellement peuplé de manière incrémentale, dans une approche ascendante (*bottom-up*) : chaque lemme ou entité complexe rencontré dans un nouveau texte à baliser, qui ne se trouve pas dans le lexique CoBRA, fait l'objet d'une nouvelle entrée où sont encodées les données morpho-syntaxiques de cet item. Par contre, comme nous l'avons mentionné plus haut, les informations morpho-syntaxiques des lexiques anglais et français ont été extraites de manière automatisée (*top down*). À l'exception de certaines entrées complexes (telles que, par exemple, des locutions nominales ou verbales), il n'y a donc pas lieu d'encoder de nouvelles entrées en anglais ni en français.

De même, les lexiques bilingues (*anglais-français* et *néerlandais-français*) ont été construits initialement dans une approche ascendante (*bottom-up*) : chaque lemme ou entrée complexe de la langue étrangère rencontré pour la première fois dans un texte à baliser fait l'objet d'un nouvel emploi (ou traduction) qui est fonction de son contexte de lecture. Les inconvénients d'une telle approche sont évidents : notre lexique comporte des trous, que sont les entrées manquantes (néerlandais) ou les entrées non traduites – ou plus exactement non garnies d'un emploi – (anglais) et qui correspondent aux mots qui n'ont pas fait l'objet d'un balisage jusqu'à présent.

Pour répondre à ce souci de cohérence interne, nous avons complété le lexique de manière systématique (par une approche descendante ou *top-down*) en y reprenant les entrées avec leur principaux emplois (traductions) selon plusieurs niveaux de fréquence que nous renseignent (i) les principaux dictionnaires explicatifs en anglais pour apprenants[7] ainsi (ii) qu'un ensemble de manuels de vocabulaire néerlandais destinés

[7] CAMBRIDGE, *Cambridge Learner's Dictionary*, Cambridge, Cambridge University Press, 2008. COLLINS, *Collins Cobuild-Advanced Learner's Dictionary*, 2009. LONGMAN, *Longman Dictionary of Contemporary English*, Longman, 2011. MACMILLAN, *Macmillan English Dictionary for Advanced Learners*, Macmillan Education, 2010. OXFORD, *Oxford Advanced Learner's Dictionary*, Oxford, Oxford University Press, 2010.

aux apprenants francophones, qui font référence dans ce domaine[8]. Pour l'anglais, nous avons également ajouté au lexique CoBRA les entrées manquantes extraites de l'*Academic Word List*[9] qui reprend les 3.000 termes à coloration académique les plus fréquents, qui sont destinés aux apprenants en milieu universitaire. Ainsi, le lexique *anglais-français* comprend actuellement 22.221 entrées (16.253 lemmes et 5.968 entrées complexes), et le lexique *néerlandais-français* 25.505 entrées (22.544 lemmes et 2.961 entrées complexes).

Nous procédons actuellement à l'étiquetage du vocabulaire de CoBRA selon les niveaux du *Cadre européen commun de référence pour les langues* – CECR[10], dans une approche qui prend en compte les acquis de l'apprenant[11]. La valeur ajoutée de ce type d'information permettra, au terme de ce travail en cours, d'associer à chaque texte balisé un niveau du CECR et de valider l'adéquation de ce texte au regard d'un niveau de cours donné, moyennant la mise en place d'une telle fonctionnalité de validation.

§ 2. Outil de lecture CoBRA

L'outil CoBRA propose un texte qui a été préalablement balisé par l'enseignant, ce qui permet à l'apprenant d'interroger chaque mot d'un clic de souris (Fig. 1, *infra*). Lorsque l'apprenant clique sur un mot du texte pour en solliciter la traduction, un tableau apparaît en bas de l'écran, avec les informations suivantes : la ligne supérieure affiche les informations grammaticales du mot : forme lemmatisée, catégorie et sous-catégorie syntaxique et formes fléchies pertinentes d'un point de vue didactique (pluriel des noms, forme comparative et superlative des adjectifs et adverbes, formes prétérit et participe passé des verbes) ; les cellules inférieures du tableau proposent plusieurs exemples du mot cliqué dans le sens du contexte de lecture (cellules de gauche) avec en regard la traduction en

[8] L. Dieltjens, M.T. Claes, J. Vanparys, P. Alkema, J. Lodewick et L. Baten, *Néerlandais Vocabulaire en contexte*, partie 1, *Woorden in context*, deel 1 (débutant), Louvain-la-Neuve, De Boeck, 2010. L. Dieltjens, M.T. Claes, J. Vanparys, P. Alkema, J. Lodewick et L. Baten, *Néerlandais Vocabulaire en contexte*, partie 2, *Woorden in context*, deel 2 (intermédiaire avancé), Louvain-la-Neuve, De Boeck, 2012.

[9] A. Coxhead, « A New Academic Word List », TESOL Quarterly, 2000, vol. 34, n° 2.

[10] Conseil de l'Europe, Conseil de la Coopération culturelle, Comité de l'éducation, Division des langues vivantes : « Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer », Paris, Didier, 2000.

[11] C. Marello, « Word lists in Reference Level Descriptions of CEFR (Common European Framework of Reference for Languages) », in R. Fjeld et J. Torjusen (eds), *Euralex 2012 Proceedings*, Oslo, University of Oslo, 2012.

français (cellules de droite) : ces exemples (concordances dans notre terminologie) peuvent illustrer différentes formes fléchies du mot et sont extraits de grands corpus bilingues alignés. Un cadre à droite reprend les entrées du glossaire personnel du texte, qui est construit par l'apprenant.

Les entités balisées peuvent être des mots simples ou composés, ainsi que des unités lexicales constituées de plusieurs mots (formes disjointes de verbes séparables néerlandais, *phrasal verbs* anglais, locutions, syntagmes verbaux ou nominaux, etc.). On notera que les exemples de traduction de chaque mot sont toujours fournis en fonction du contexte de lecture, les ambiguïtés ayant été levées lors du balisage préalable du texte.

Les concordances qui illustrent chaque traduction sont affichées sur des fonds de couleurs différentes[12] en fonction du domaine de spécialité de leur provenance (économie et finance, informatique, droit, sciences et techniques, politique, art et histoire). Ces concordances, dont le nombre total s'élève à près de onze millions, sont générées et filtrées automatiquement, et ne reprennent, à l'intérieur de chaque corpus (ou *couleur*) de spécialité, qu'un seul exemple de chaque traduction (voire de chaque synonyme qui y est associé). Ce filtrage évite de surcharger le nombre de concordances qui illustrent l'emploi d'un mot ou d'une expression par souci de lisibilité.

L'affichage des concordances étant limité à une fenêtre de taille arbitraire, l'apprenant ne peut donc qu'en lire un extrait dans certains cas. Un clic sur le mot en langue étrangère dans son contexte (cellule de gauche) fait apparaître dans la même page le contexte complet du mot retenu (phrase, voire paragraphe en langue étrangère avec sa traduction en français) avec les références du corpus d'illustration (Fig. 1).

Une fonction (intitulée *mon glossaire*) permet à l'étudiant de se constituer un glossaire à la carte, à partir des mots qu'il a préalablement interrogés (par un clic) dans un ou plusieurs textes de leçon, en les ajoutant à sa propre liste de vocabulaire par un clic supplémentaire sur l'icône à proximité du mot, qui est dédiée à cette fonctionnalité (Fig. 1). Ce glossaire personnalisé peut être mis à jour à tout instant par l'étudiant qui souhaite en supprimer un ou plusieurs termes.

Enfin, le glossaire de l'étudiant peut être téléchargé en format imprimable (.xls) et compatible avec des outils tels que *Quizlet*[13], où l'étudiant parcourt et révise le vocabulaire sous forme de cartes ou d'exercices ludiques.

---

[12] Il n'est pas possible de rendre un tel affichage polychrome dans le cadre du présent article.

[13] QUIZLET, 2022, https://quizlet.com.

Outre la traduction d'un mot ou d'une expression proposée selon son contexte de lecture, le glossaire personnel de l'étudiant constitue une des valeurs ajoutées cardinales de l'outil CoBRA, en ce qu'il donne à l'étudiant seul un rôle central dans l'élaboration d'un support de vocabulaire selon ses besoins.

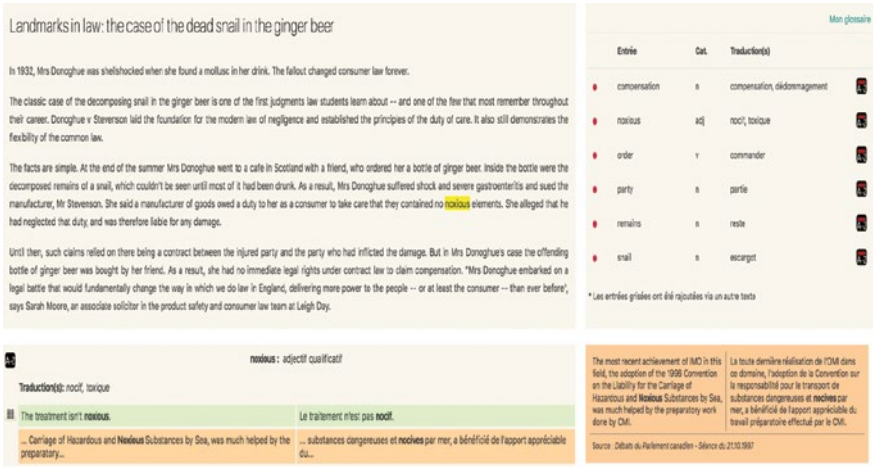

Figure 1 – Outil de lecture CoBRA (mot *noxious* activé)

L'outil de lecture CoBRA permet à l'enseignant de paramétrer l'affichage des informations visibles révélées par un clic sur chaque mot ou expression. Ainsi, les formes fléchies d'un item peuvent être affichées en tout ou en partie, la traduction de cet item peut apparaître systématiquement en complément des concordances, ou uniquement en l'absence de celles-ci, des annotations (définitions et commentaires) liées à une entrée peuvent être proposées, les concordances peuvent être affichées en mode bilingue (langue source et langue cible) ou unilingue (langue source uniquement).

Pour les besoins d'un cours spécifique, les concordances liées à chaque mot ou expression d'un texte de ce cours peuvent être ordonnées par spécialité, et les corpus d'illustration peuvent aussi faire l'objet d'une sélection par spécialité de la part de l'enseignant. Ainsi, un texte en néerlandais à l'attention d'étudiants en droit, par exemple, privilégiera les concordances de cette spécialité en plaçant celles-ci en tête du tableau pour illustrer les termes de ce texte, et n'affichera pas les concordances issues des corpus d'informatique ou d'histoire.

L'enseignant peut également générer une liste bilingue (*glossaire* dans notre terminologie) de tous les mots et expressions avec leurs propriétés grammaticales, liés à un texte ou à une collection. Notons aussi qu'une

fonction du module d'administration de CoBRA permet également de calibrer tout nouveau texte (non balisé) en référence à un glossaire existant, c'est-à-dire d'en calculer la *couverture lexicale* sur la base du lexique d'une ou plusieurs *collections* (partielles ou complètes) de son choix.

Par ailleurs, un dictionnaire bilingue (*néerlandais-français* et *anglais-français*) est proposé à l'apprenant, indépendamment du module de lecture décrit ci-dessus (Fig. 2). Ce dictionnaire non contextualisé, qui est également accessible sans restriction en dehors de la plateforme d'apprentissage, reprend les informations suivantes associées à chaque entrée (mot ou expression) : (i) forme lemmatisée, (ii) catégorie et sous-catégorie syntaxique et (iii) formes fléchies pertinentes d'un point de vue didactique (pluriel des noms, forme comparative et superlative des adjectifs, formes prétérit et participe passé des verbes), (iv) tous les emplois (sens) liés à cette entrée et (v), dans le cas d'une entrée simple (lemme), toutes les entrées complexes liées à cette entrée. Ici aussi, chaque emploi est illustré par des concordances issues de très grands corpus bilingues alignés, recouvrant la langue usuelle et les domaines de spécialité décrits plus haut.

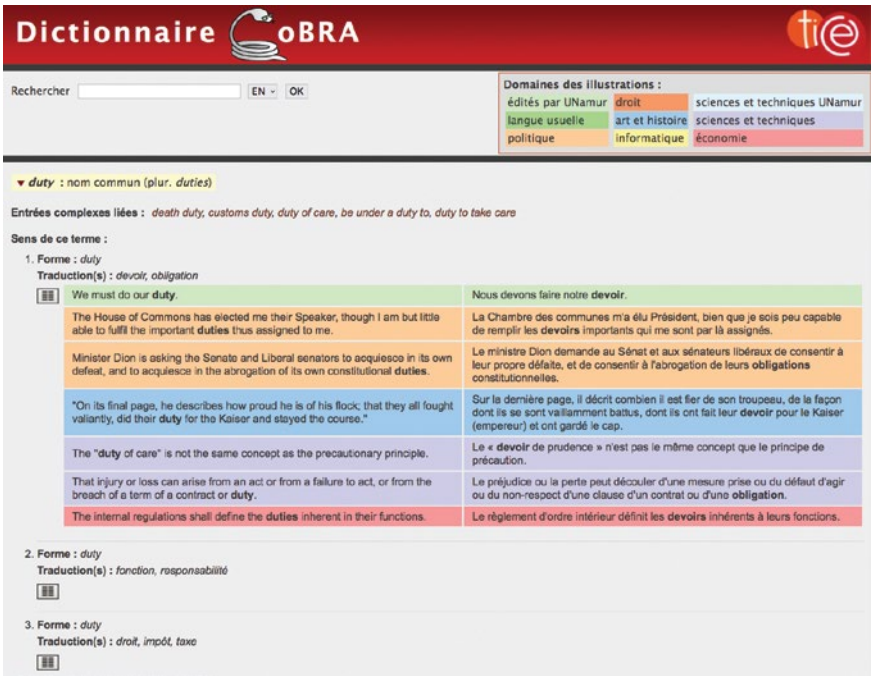

Figure 2 – Dictionnaire CoBRA à l'usage de l'apprenant avec l'entrée *duty* (les concordances illustrant les sens 2 et 3 de cette entrée n'ont pas été affichées par manque d'espace)

§ 3. Module de balisage CoBRA

Le module de balisage CoBRA permet de combiner les différentes étapes menant à l'étiquetage d'un texte de leçon, où, pour chaque mot ou expression, on détermine de manière semi-automatique l'entrée correspondante dans un lexique *langue étrangère-français*. Les étapes d'étiquetage s'effectuent itérativement sur chaque paragraphe du texte, ce qui permet à l'utilisateur d'interrompre au besoin le balisage d'un texte au terme d'un de ses paragraphes, et de reprendre celui-ci sans dommage par la suite.

A) Repérage des entités manquantes

La première étape du balisage d'un texte consiste en un repérage des entités complexes (*locutions* et *verbes complexes*) à la fois présentes dans le texte et dans la base de données lexicale (Fig. 3). À ce stade, l'enseignant peut compléter le texte par les expressions manquantes qu'il a repérées, en encodant celles-ci dans le lexique.

Entrées complexes détectées

Until then, such claims relied on there being a contract between the injured party and the party who had inflicted the damage. But in Mrs Donoghue's case the offending bottle of ginger beer was bought by her friend. As a result, she had no immediate legal rights under contract law to claim compensation. "Mrs Donoghue embarked on a legal battle that would fundamentally change the way in which we do law in England, delivering more power to the people – or at least the consumer – than ever before," says Sarah Moore, an associate solicitor in the product safety and consumer law team at Leigh Day.

expressions détectées : digital output (n) – as a result (conj) – there are (v) – there was (v) – embark on (prepositional verb) – there is (v) – at least (adv) – rely on (prepositionla verb) – do in (phrasal verb)

Figure 3 – Repérage des constituants multiples (*expressions*)

On pourra alors procéder à l'étape ultérieure du balisage proprement dit, qui retournera le texte où chaque expression identifiée par le moteur est présentée sous la forme d'un menu déroulant permettant à l'enseignant de confirmer ou non la prise en compte de cette expression (Fig. 5).

Lemmes détectés

Landmarks in law: the case of the dead snail in the ginger beer

In 1932, Mrs Donoghue was shellshocked when she found a **mollusc** in her drink. The fallout changed consumer law forever.

The classic case of the decomposing snail in the ginger beer is one of the first judgments law students learn about – and one of the few that most remember throughout their career. Donoghue v Stevenson laid the foundation for the modern law of negligence and established the principles of the duty of care. It also still demonstrates the flexibility of the common law.

The facts are simple. At the end of the summer Mrs Donoghue went to a cafe in Scotland with a friend, who ordered her a bottle of ginger beer. Inside the bottle were the decomposed remains of a snail, which couldn't be seen until most of it had been drunk. As a result, Mrs Donoghue suffered shock and severe **gastroenteritis** and sued the manufacturer, Mr Stevenson. She said a manufacturer of goods owed a duty to her as a consumer to take care that they contained no noxious elements. She alleged that he had neglected that duty, and was therefore liable for any damage.

Until then, such claims relied on there being a contract between the injured party and the party who had inflicted the damage. But in Mrs Donoghue's case the offending bottle of ginger beer was bought by her friend. As a result, she had no immediate legal rights under contract law to claim compensation. "Mrs Donoghue embarked on a legal battle that would fundamentally change the way in which we do law in England, delivering more power to the people – or at least the consumer – than ever before," says Sarah Moore, an associate solicitor in the product safety and consumer law team at Leigh Day.

Couverture

- lemmes détectés avec traduction : 288/306 (94,12 %)
- **lemmes sans traduction : 2/306 (0,65 %)**
- lemmes absents du dictionnaire : 16/306 (5,23 %)

Figure 4 – Repérage des mots (lemmes) absents du lexique ou dépourvus d'emploi.

Cette étape se poursuit avec l'identification de chaque mot (*lemme* dans notre terminologie) à la fois présent dans le texte et dans la base de données lexicale, en y associant les entrées dont ce mot constitue une flexion, ainsi que toutes les traductions déjà encodées pour ces entrées. Les mots absents du lexique et les mots dépourvus de traduction sont identifiés par un code de couleur spécifique que nous avons converti ici en diverses polices de caractères par souci de lisibilité (Fig. 4).

Une fois les mots manquants et les traductions manquantes encodés dans le lexique, l'utilisateur peut procéder au balisage du texte proprement dit, à l'aide de menus déroulants qui permettront de lever les éventuelles ambiguïtés dues à des phénomènes de polysémie et d'homonymie, et de fournir une traduction contextualisée des mots de ce texte (Fig. 8).

B) Balisage d'un texte

Nous avons vu que l'outil de balisage CoBRA assiste l'enseignant dans la tâche cruciale de désambiguïsation des termes et expressions d'un texte. Plus précisément, pour chaque expression ou terme qui revêt plusieurs sens, la version actuelle de CoBRA propose une série de choix à valider, selon une heuristique qui prend notamment en compte les paramètres suivants, que nous allons détailler : la nature des items (*lemme* vs. *entrée complexe*), leur catégorie syntaxique, la taille de la fenêtre courante du texte à analyser et la fréquence d'utilisation des différents sens d'une entrée lexicale.

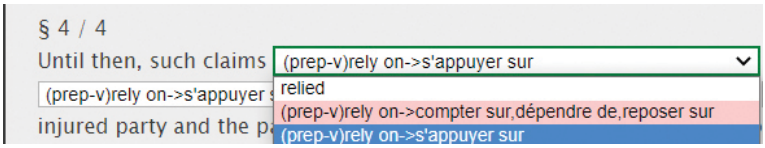

Figure 5 – Exemple de balisage des expressions (balisage de *rely on*)

Nous avons indiqué que l'outil propose d'abord le balisage des constituants multiples avant de passer au balisage des lemmes. Ces deux phases distinctes permettent de filtrer les constituants qui composent une expression en proposant à l'enseignant d'identifier ces constituants comme tels, ce qui limite ainsi la liste des traductions possibles pour chacun de ces constituants lors de la phase ultérieure de balisage des lemmes.

L'étape de balisage semi-automatique des *lemmes* au sens strict (items constitués d'une seule chaîne de caractères) demeure le goulet d'étranglement dans la production des ressources par CoBRA. En particulier, la procédure de désambiguïsation des termes peut s'avérer coûteuse en temps. Aussi, nous avons optimisé cette procédure à l'aide d'une heuristique dont nous allons brièvement décrire les contraintes. Une telle heuristique laissera toutefois à l'utilisateur le choix final de la balise appropriée lors de la phase de désambiguïsation.

Une première contrainte consiste à prendre en compte la catégorie syntaxique d'un terme-candidat lors du balisage de celui-ci. Cette information qui a été obtenue préalablement par *TreeTagger*, un outil d'étiquetage syntaxique intégré à CoBRA[14], permet ainsi de lever automatiquement certaines ambiguïtés lors du calcul final de la traduction[15].

[14] H. SCHMID, « Probabilistic Part-of-Speech Tagging Using Decision Trees », Proceedings of International Conference on New Methods in Language Processing, Manchester, United Kingdom, 1994.

[15] Les résultats d'un tel étiquetage automatique ne sont pas exempts d'erreurs : la catégorie syntaxique renvoyée par un tel outil nécessite donc une validation *a posteriori* de la part de l'utilisateur.

Une seconde contrainte de l'outil de balisage consiste à ordonner les emplois d'un terme polysémique selon leur fréquence dans la collection des textes déjà balisés à laquelle appartient le texte-candidat à baliser, chaque collection étant caractérisée par un type spécifique (*juridique*, *économique*, *informatique*…). L'hypothèse étant que le sens privilégié d'un mot polysémique est identique dans tout texte du même type. Si le terme à désambiguïser n'est pas repris dans une collection donnée, ses emplois sont ordonnés selon leur fréquence dans l'ensemble des textes déjà balisés de toutes les collections. L'hypothèse étant que le sens privilégié d'un mot polysémique est son sens le plus fréquent.

Au cas où la balise affectée à un item a été validée par l'utilisateur dans le premier paragraphe d'un texte, la procédure de désambiguïsation proposera par héritage la même balise à toutes les occurrences de ce terme dans le reste du texte (en caractères gras dans le menu déroulant décrit plus haut), sans éliminer toutefois de la liste les autres emplois éventuellement associés à ce terme. L'hypothèse étant que le sens privilégié d'un mot polysémique est identique au sein d'un même texte.

Ainsi, le balisage des lemmes distinguera quatre cas de figure qui répondent à la combinaison de ces contraintes (Fig. 6) :

Balisage – Traitement des lemmes – Aide

The classic **case** of the **decomposing** snail **in** the **ginger beer is one** of the first judgments **law** students learn about – and **one** of the few **that most** remember **throughout their** career. Donoghue v Stevenson **laid** the **foundation for** the modern **law** of **negligence** and established the principles of the *duty of care*. It also **still demonstrates** the flexibility of the *common law*.

Encoder et continuer Encoder et quitter Quitter

- *élément déjà étiqueté*
- mot étiqueté automatiquement
- mot à valider (emploi unique dans la catégorie syntaxique probable)
- **mot à valider** (emploi le plus fréquent dans la catégorie syntaxique probable)

Figure 6 – Exemple de balisage des lemmes (avant validation et désambiguïsation)

(i) éléments déjà étiquetés (*italiques*) : il s'agit de mots qui ont été préalablement repris au sein d'une expression déjà balisée. De tels mots ne font donc plus l'objet d'une balisage spécifique ;

(ii) mots étiquetés automatiquement : il s'agit de lemmes avec une catégorie syntaxique unique et un emploi unique dans notre dictionnaire. De tels lemmes sont exempts d'ambiguïté syntaxique et sémantique, et se voient donc automatiquent attribuer la balise qui renvoie à l'emploi unique lié à cette entrée ;

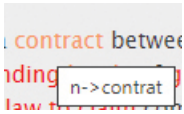

Figure 7 – Désambiguïsation (syntaxique) automatique du mot *contract*

(iii) mots monosémiques à valider (surlignés) : il s'agit de lemmes avec plusieurs entrées (catégories syntaxiques) dans notre dictionnaire dont l'analyseur syntaxique a filtré la catégorie syntaxique probable. Dans cette catégorie, ces lemmes ont un emploi unique dans notre dictionnaire. En passant la souris sur le mot-candidat, la balise proposée automatiquement apparaît en *pop-up*. Si la balise proposée est correcte, le mot a été étiqueté correctement et automatiquement (Fig. 7). Si la balise est incorrecte en raison du comportement erroné de l'analyseur syntaxique (Fig. 8), l'utilisateur peut activer un menu déroulant pour étiqueter le mot à l'aide de la balise adéquate (Fig. 9.) ;

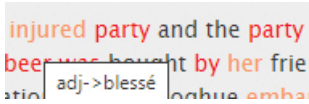

Figure 8 – Désambiguïsation automatique erronnée du mot *injured*

(iv) mots polysémique à valider (**gras**) : il s'agit de lemmes avec plusieurs entrées (catégories syntaxiques) dans notre dictionnaire, dont l'analyseur syntaxique a filtré la catégorie syntaxique probable, et dont l'outil de balisage a retenu l'emploi le plus fréquent dans cette catégorie (dans la *collection* courante, ou à défaut dans tous les textes des *collections* existantes). Ici aussi, la balise proposée automatiquement apparaît en *pop-up* lors du passage de la souris. Si la balise est correcte, le mot a été étiqueté correctement et automatiquement. Si la balise est incorrecte (comportement erroné de l'analyseur syntaxique et/ou emploi erroné dans le contexte courant), l'utilisateur peut activer un menu déroulant pour étiqueter le mot selon l'emploi adéquat.

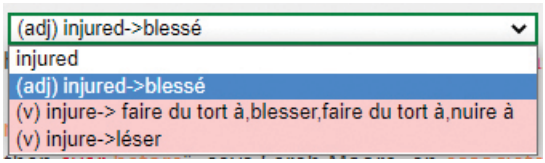

Figure 9 – Désambiguïsation manuelle du mot *injured* par menu déroulant

SECTION 3. – Bilan

L'outil CoBRA a fait l'objet de plusieurs évaluations de la part des apprenants et des enseignants, dont nous reprendrons ici les éléments les plus saillants. Ainsi, une enquête a été menée auprès de 325 apprenants francophones de l'enseignement universitaire dans le cadre d'un cours de néerlandais juridique de 2e année de bachelier en droit[16]. L'objectif de ce cours est de rendre l'étudiant capable de lire et comprendre des documents rédigés en néerlandais, qui traitent de thèmes généraux relevant du droit public, du droit judiciaire et du droit privé, ainsi que des textes plus techniques de jurisprudence et de doctrine. C'est donc une connaissance réceptive de la langue qui est visée et évaluée.

Cette étude indique que l'utilité de l'outil CoBRA (dans sa version de 2004, exempte de glossaire personnel) dans un tel contexte d'apprentissage spécialisé est très nettement perçue (par près de 90 % des apprenants). Nous y avons également observé que CoBRA est considéré par une majorité d'étudiants comme un moyen d'auto-apprentissage en préparation du cours (77 %) et une aide lors de l'exposé (82,4 %).

Par ailleurs, depuis l'année académique 2011-2012, CoBRA est mis à la disposition de plus de 5.000 étudiants du bachelier de l'UNamur et de UCLouvain, pour l'apprentissage de l'anglais et du néerlandais, dans des niveaux et des contextes d'apprentissage différenciés (droit, histoire, informatique, sciences économiques, sociales et de gestion, sciences politiques et de communication, sciences, médecine, pharmacie, psychologie).

Parallèlement, CoBRA est utilisé par le Forem depuis 2012 pour la formation en néerlandais et en anglais des apprenants issus du monde du travail, notamment dans les secteurs du *e-business* et des technologies de l'information.

[16] G. Deville et L. Dumortier, « Évaluation d'un outil en ligne d'aide à la lecture de textes en langue étrangère », in Actes de la journée d'étude de l'ATALA TAL et Apprentissage des langues, Grenoble, 2004, pp. 93-102.

Les premières évaluations de l'outil dans les cours d'anglais et de néerlandais de 1re année en baccalauréat de droit à l'UCLouvain indiquent que les 310 étudiants interrogés s'approprient d'autant mieux un tel outil d'aide à la lecture, que l'enseignant accompagne ceux-ci en début d'année dans la prise en main de CoBRA, et dans son exploitation optimale pour leur apprentissage.

D'autre part, une enquête a été menée en 2011-2012 auprès de 73 étudiants de l'enseignement universitaire de 3e année du bachelier en droit dans le cadre d'un cours de droit comparé et de terminologie juridique anglaise à l'UNamur. L'objectif de ce cours est de rendre l'étudiant capable de lire, comprendre, et commenter (en français) des textes juridiques (décisions de jurisprudence) rédigés en anglais.

L'étude révèle que 78 % des étudiants interrogés sont (plutôt ou tout à fait) d'avis que l'outil CoBRA facilite la compréhension des décisions de jurisprudence. 90,4 % des étudiants interrogés sont (plutôt ou tout à fait) d'avis que l'outil facilite également la préparation du cours donné en classe (en français) par l'enseignant. De l'avis du titulaire du cours, les commentaires associés à certaines entrées du lexique (qui remplacent utilement une traduction inexistante) ont rendu les étudiants plus sensibles aux limites d'une traduction littérale du langage juridique.

En conclusion, CoBRA est un outil informatique d'aide à l'apprentissage des langues qui ne remplace pas l'enseignant dans sa classe ; il est complémentaire au travail de ce dernier. CoBRA permet à l'apprenant d'être acteur de sa formation : celui-ci dispose seul des clés d'accès au vocabulaire qu'il estime nécessaire à sa compréhension d'un texte écrit.

L'approche que nous avons adoptée dans la création des ressources d'un tel outil, et en particulier dans le balisage de textes, vise à fournir une aide automatisée à l'enseignant, à qui reviennent *in fine* les choix appropriés. Cette approche s'inscrit dans le traitement interactif de la langue (TIL), qui est une voie de recherche prometteuse explorée par M. Zock[17]. Nous avons affaire à une machine (l'outil CoBRA) qui permet à l'être humain (l'enseignant) de passer d'une solution partielle (les connaissances linguistiques mises à sa disposition) à la solution complète (le choix d'un étiquetage lexical adéquat).

[17] M. Zock et G. Lapalme, *Du TAL au TIL*, Actes de la conférence Traitement automatique des langues naturelles (TALN), Montréal, 2010.