Consignes :

* Créez la table des matières du document qui suit en vous appuyant sur la table des matières donnée en PDF.
* Personnalisez le style des titres niveau 3 (couleur bleue, gras, italique et souligné).

**Nom :**

**Prénom :**

**Groupe :**

Introduction

Pour le thème de ce mémoire de licence nous avons choisi la linguistique de corpus, sa présentation en abrégé suivie de son support technique.

La linguistique de corpus est la branche de la linguistique qui exploite les technologies informatiques pour sa recherche. Elle participe à la création de dictionnaires, de grammaires etc., elle facilite aussi l’analysation linguistique de divers phénomènes. Tous ces « aptitudes » la font indispensable dans le domaine des sciences humaines. Auparavant, les compilateurs rassemblaient manuellement des mots et des exemples de leur emploi et successivement ils les notaient, ce qui était très fatigant et prenait beaucoup de temps. Dans cette problématique, un grand progrès y est représenté par la naissance des premiers corpus et par leur transformation sous forme électronique plus tard. Cette informatisation a aussi rendu possible d’analyser divers faits ou problèmes que l’utilisateur peut ensuite évaluer statistiquement et ainsi de suite. La taille de certains corpus et leur équilibrage assurent que les données obtenues d’un tel corpus peuvent être considérées comme les représentatives. Les corpus ne sont pas cependant destinés seulement aux linguistes et aux lexicographes, leur champ d’action est beaucoup plus vaste – les sociologues et les sociolinguistes, les traducteurs, les spécialistes dans le domaine des médias et de la communication de masse, les psychologues et d’autres les aussi utiliseront pour leur travail.

Aperçu de la linguistique de corpus

Définition du corpus linguistique

D’abord, il est nécessaire d’expliquer ce que « le corpus » représente en linguistique. Ce mot descend de la langue latine où il désigne « corps », « chair du corps », « personne, individu », « corps inanimé, cadavre» ou « corps, ensemble, tout». En français aussi, il a plusieurs sens. Dans la langue de droit, il exprime « recueil de pièces, de documents concernant une même discipline » . Dans la terminologie linguistique, il est décrit comme « ensemble fini d’énoncés réels réuni en vue de l’étude d’un phénomène linguistique ». Dans cet égard, « c’est donc une base matérielle qui sert à une analyse linguistique et à une description d’une langue écrite aussi bien qu’une langue parlée (transcrite pour des besoins de recherche) ».

Histoire brève de la linguistique de corpus

L’histoire de ce domaine de la linguistique est liée surtout aux pays anglophones. Et même si elle se développe aussi dans d’autres pays à présent, le centre de la recherche reste toujours là-bas.

La linguistique de corpus travaillant avec les corpus sous la forme électronique est née dans la deuxième moitié du XXe siècle, plus précisément dans les années soixante, mais cela ne signifie pas encore qu’avant, les linguistes ne se soient intéressés ni à l’utilisation authentique de la langue ni aux citations – ces corpus non-électroniques étaient cependant connus sous les noms « fichier », « archives » et ainsi de suitte. Comme exemple, on peut mentionner le Dictionary of English Language publié au début du XVIIe siècle sous la direction de Samuel Johnson, l’American Dictionary of the English Language de l’année 1828, l’English Dialect Dictionary de J. Wright du tournant du VIIIe et du XIXe siècle, l’Oxford English Dictionary dont le dernier tome a paru en 1928. Dans la première moitié du XXe siècle, c’est la pédagogie linguistique qui commence à manipuler plus intensivement les corpus, par ex. le Teacher’s Wordbook of 30.000 Words d’E. L. Thorndike et I. Lorge. Des grammaires aussi s’appuient sur beaucoup de matériaux linguistiques et c’est pourquoi elles, dès la deuxième moitié du XXe siècle avant tout, profitent des collections de citations. Le corpus non-électronique le plus remarquable destiné à ce but était le Survey of English Usage Corpus dont le projet a été ouvert par Randolph Quirk en 1959. Son étiquetage a inspiré des créateurs du Brown Corpus. Le SEU Corpus faisait aussi la base de la grammaire anglaise le Comprehensive Grammar of the English Language et dans la deuxième moitié des années soixante-dix, il a été transformé en forme électronique et ainsi sa partie parlée fait maintenant partie intégrante du London-Lund Corpus.

fermé comprenant cent millions de formes lexicales. Il se divise en deux parties – 90 % de textes font une partie écrite et 10 % représente une partie parlée.

Division des corpus linguistiques

 L’utilisation des corpus peut être diverse et de ce fait il découle division selon différents aspects. Il est possible de juger des critères comme par exemple de l’étendue du corpus, son intervalle de temps, de la sélection de textes qui deviendront sa partie composante, la langue ou les langues de ces textes.

 L’une de divisions possibles est la division en corpus d’une langue, généralement d’une langue nationale, et en corpus parallèles. Cette problématique est mentionnée plus en détail ci-dessous. Une autre distingue les corpus diachroniques d’avec des corpus synchroniques.

Corpus diachroniques et synchroniques

Dans le cas de ces deux types de corpus, c’est l’intervalle de temps de textes compris qui y représente un critère distinctif.

Corpus diachroniques

Les corpus diachroniques comportent les textes d’une période de temps plus longue que les corpus synchroniques et ils sont fermés. L’Helsinki Corpus of English Texts : Diachronic and Dialectal, le Lampeter Corpus ou le Diakorp peuvent servir d’exemple de ce type-là.

Corpus synchroniques

Au contraire, les corpus synchroniques ont pour leur but de ne comprendre qu’une relativement courte période de temps, ils n’enregistrent pas donc l’évolution de la langue et des changements en elle. Ils sont construits en considération de leur représentativité – alors il est veillé à la représentation équilibrée des différents genres. Dans la majorité des cas, il s’agit de corpus de la langue actuelle (lors de la naissance du corpus concret).

Quelques corpus synchroniques – ceux qui n’ont pas été créés avec un but spécifique - ne préfèrent aucun domaine de la linguistique, aucune niveau de la langue ou appartenance de style et c’est pourquoi ils sont comptés parmi les corpus généraux. Un autre attribut de ce type de corpus est son équilibrage. Cela signifie qu’en formant le corpus, on prend en considération plusieurs aspects : différents types de genres littéraires (le style journalistique y compris - les corpus généraux travaillent avec la presse quotidienne etc.), médias de transmission (la langue parlée et écrite) et des registres de langue de textes choisis (matériaux d’un caractère privé ou public, officiel).

 .

Corpus nationaux et parallèles

Étant donné que les différents corpus peuvent être souvent classés en plusieurs sortes de corpus, il s’y trouve aussi les corpus spécialisés disposant de certains traits de corpus nationaux dans ce sous-chapitre.

Corpus nationaux

Le ČNK est le projet universitaire de l’ÚČNK à la Faculté des Lettres de l’Université Charles de Prague visant une création d’un corpus informatisé, surtout d’une langue écrite. Ce corpus date de l’année 1994 et il se divise en deux parties : la partie synchronique et diachronique, ce qui est déjà mentionné ci-dessus.

La partie synchronique est de plus divisée en corpus écrits et en corpus parlés. Le premier groupe est formé des sous-corpus suivants : ORWELL, LITERA, SYNEK, KSK - DOPISY, FSC2000, SYN2000, SYN2005 et SYN2006PUB.

L’ORWELL est basé sur le texte du roman 1984 de George Orwell. Il comprend à peu près 80 mille mots - grâce à son étendue petite relativement, qui a rendu possible de contrôler l’analyse morphologique faite automatiquement, il est étiqueté presque sans fautes.

Corpus parallèles

Corpus PAROLE (PARallèle, Oral, en Langue Étrangère)

Le but du projet PAROLE est de créer un assortiment des corpus et des lexiques pour tous les langues de l’Union européenne suivant les paramètres unifiés. Ces paramètres se rapportent au temps de production et à la classification de ses supports divisés en catégories « Livre », « Journal », « Périodique » et « Divers ». L’exigence sur l’harmonisation des corpus particuliers se réfère aussi à leur étiquetage morpho-syntaxique.

 Ce projet, réalisé entre les années 1996 – 1998, est basés sur la production orale – enregistrée, transcrite et annotée - d’apprenants du français, de l’anglais, de l’italien etc. étant sur divers niveaux de parole. Les divers corpus, donc construits pour chaque langue, sont formés de plus de 20 millions de mots. Il est possible de les obtenir sous forme de CD-ROM (le Corpus Français PAROLE etc.), par l’intermédiaire de l’ELRA par exemple. Ces corpus ne sont pas interconnectés mutuellement. En fait, il s’agit de corpus nationaux spécialisés créés selon les mêmes critères.

Support technique de la linguistique de corpus

Outils logiciels et explication de certains termes

D’abord, il est nécessaire de convertir les textes qui ont été choisis à faire partie du corpus. Cela signifie qu’ils doivent être modifiés en même forme. Pendant ce processus, les textes isolés gagnent un en-tête d’identification grâce auquel le linguiste peux acquérir plusieurs renseignements sur le texte dont par exemple une expression cherchée descend. À part cela, le texte brut est difficile à traiter et c’est pourquoi les corpus annotés sont mieux exploitables pour les linguistes. Le texte annoté est né au moment où le texte brut est enrichi d’annotations. On peut annoter une structure logique du texte (par exemple des chapitres, des alinéas), une apparence typographique ou des phénomènes linguistiques. Les corpus peuvent être annotés manuellement, ce qui est cependant trop exigeant, ou à la machine. La deuxième manière dite est bien plus rapide, mais aussi moins précise. Il faut dire que « dans le cadre du traitement automatique des langues naturelles, les corpus soient souvent utilisés pour entraîner et tester des modèles » et qu’on distingue deux types de corpus : les corpus étiquetés et les corpus annotés.

Gestionnaire de corpus

Les gestionnaires de corpus fonctionnent de façon similaire à celle des moteurs de recherche sur Internet, mais ceux-ci ne visent pas primordialement l’acquisition de données linguistiques. En revanche, les gestionnaires de corpus manipulent des outils divers pour la recherche et la classification de matériaux textuels. Entre autres, on peut nommer le Corpus Query Processor (CQP) ce qui est un outil créé pour trouver des renseignements d’un corpus codé par l’IMS Corpus Workbench (CWB), développé à l’Université de Stuttgart. Ce gestionnaire est l’un des outils logiciels qui sont nés dans le cadre du projet sur Text Corpora and Exploration Tools et à présent, il est utilisé aussi en Faculté d’informatique de l’Université Masaryk de Brno et dans l’Institut du ČNK. L’avantage de ce système est par exemple le fait qu’il est capable de conserver une quantité arbitraire d’attributs de chaque mot cherché (quant au tchèque, outre une forme de base d’un mot, il peut aussi garder toutes ses catégories grammaticales et ainsi de suite). Une autre fonction de ce système représente les soi-disant questions de croissance . Ces questions permettent de formuler successivement des questions complexes. En plus, il y a la possibilité de la création d’un sous-corpus en vertu des résultats des différentes questions avec lequel on peut travailler encore plus tard. Les résultats trouvés sont affichés en format KWIC.

Programmes de concordance

La fonction des programmes de concordance consiste à classer et à compter les objets trouvés dans le corpus. Dans le corpus « brut », ils sont représentés par des formes lexicales, de la ponctuation, etc. Si le corpus n’est pas complété d’une autre information, le programme de concordance ne peut pas distinguer certaines homonymies. En ce qui concerne la langue française, comme un exemple on peut citer le mot « livre » - il s’agit de l’une de représentants des homographes, « mots qui s’écrivent de la même manière mais qui ont des sens différents (un livre qui se lit vite / une livre de légumes) » . C’est la raison pourquoi les programmes qui sont les analyseurs grammaticaux en quelque sorte, font aussi partie des outils de corpus. Dans ce cas, on parle habituellement de l’étiquetage (tagging en anglais) et de programmes d’étiquette (taggers en anglais) de divers niveaux. Les programmes soit travaillent avec ces analyseurs, soit ils les presque comprennent.

 Lemmatisateur

Le lemmatisateur fait aussi partie des outils logiciels qui facilite le travail avec le corpus. Son fonction consiste à ajouter un lemme respectif à chaque occurrence lexicale dans tous les textes du corpus. Étant donné que toutes les langues ont leurs spécificités, ce programme doit être créé indépendamment pour chacune d’elles.

Langages informatiques

HTML

L’HTML est un langage de balisage prédominant pour la formation de documents hypertextes. Il rend possible d’implanter de l’hypertexte dans le contenu des pages Web, de structurer sémantiquement ces pages, de les mettre en forme etc. L’HTML donc prête les moyens pour décrire des renseignements basés sur un texte compris dans un document et ensuite pour compléter ce texte par des formes interactives, des images incluses et ainsi de suite. Il s’agit de l’application du langage de balisage universelle plus vieux, le SGML. L’HTML est le langage qui se développe – c’est soi-disant le W3C qui s’occupe de sa normalisation (il n’émet pas des normes obligatoires, mais plutôt des recommandations).

 PHP

Le PHP est un langage de script (côté serveur HTTP) duquel le but originel a été la production des pages Web dynamiques. Cette fonction du PHP reste la principale, mais il y a aussi des autres : il représente le langage de programmation en ligne de commande et il sait écrire des applications clientes graphique, même s’il ne fait pas partie des meilleurs langages qui disposent de cette aptitude. Si l’utilisateur veut exploiter le PHP à son objet primaire, il ne se passe pas de trois composants : un analyseur PHP, un serveur Web et un navigateur Web. En revanche, ni serveur Web ni navigateur Web ne sont indispensable à l’écriture des scripts PHP. Pour cette activité, il suffit d’avoir à la disposition l’exécutable PHP. En ce qui concerne les applications clientes graphiques, on peut les créer à l’aide du PHP-GTK ce qui est l’extension du PHP utilisant la bibliothèque GTK+. Le langage PHP est d’accès libre, il s’agit d’un gratuiciel (freeware en anglais). Les scripts PHP peuvent être intégrés aux pages HTML.

Utilisation du langage PHP à la linguistique de corpus

Comme il a été dit dans le chapitre précédent, le langage PHP est souvent incorporé directement dans la structure de documents en langage HTML (le nom d’un pareil fichier est fini par <php>). Dans la majorité des cas, les scripts PHP sont cependant effectués du côté serveur où il arrive à leur interprétation via un interpréteur de script. L’utilisateur donc ne voit que le résultat de leur activité qui a été envoyé au navigateur. Il est donc impossible d‘afficher l’enregistrement du PHP, dans le code source du page, parce que le PHP reste toujours sur le serveur. Un seul code que nous pouvons visualiser est celui du HTML.

Préparation de textes correspondants

Pour ce mémoire, nous avons utilisé l’œuvre Stupeur et tremblements de l’écrivain belge contemporain Amélie Nothomb.

D’abord, il faut obtenir deux versions de ce livre – l’original et la traduction. L’une devrait se rapporter à l’autre, cela vaut dire que la traduction gagnée devrait être la traduction de l’édition de l’original dont nous disposons.

Le pas suivant est représenté par la conversion des textes en même forme, en fichier à l’extension de nom <rtf> dans le cas idéal. En principe, nous pouvons l’atteindre par deux voies – soit par l’acquisition des œuvres sous forme digitalisée (par ex. de la maison d’édition), soit par la numérisation des livres. La première façon est sans doute plus commode et plus rapide, mais la deuxième est plus certaine – la maison d’édition n’a pas à posséder le fichier exigé et ainsi de suite.

Démonstration du fonctionnement du PHP

Pour pouvoir de montrer comment le script PHP mentionné plus haut fonctionne, nous avons décidé de citer la préposition « à » comme un exemple. Tout d’abord nous nous occuperons de sa position et fonction dans la phrase français et tchèque et puis, nous passerons à la description des différents cas.

Prépositions

Les prépositions sont des mots ou des locutions invariables. Elles servent à construire – c’est-à-dire à rattacher l’un à l’autre – deux termes ou deux groupes non parallèle, c’est-à-dire qui n’assument pas la même fonction.

Dans la Grammaire du français classique et moderne, les prépositions sont désignées comme « les struments » – par le mot de Damourette et Pichon – ce qui veut dire « qu’elle explicitent le rapport syntaxique de deux termes qui, eux, assument une fonction ».

À cela se rapporte « le régime » de la préposition. Il est représenté le plus souvent par un complément circonstanciel, un objet indirect, un complément du verbe passif, un complément déterminatif, un complément d’adjectif ou d’adverbe. En conformité de l’usage ordinaire, le régime suit immédiatement la préposition avec laquelle il forme souvent une unité sémantique.

Préposition « à »

Étant donné que chaque œuvre qui s’occupe de la grammaire française, offre une diverse vue sur la division de l’emploi de la préposition « à », nous avons décidé de montrer son rôle dans la phrase à la base de Francouzská mluvnice (pp. 484 - 488) écrite par Josef Hendrich, Otomar Radina et Jaromír Tláskal. Les exemples mentionnés sont choisis des premières cinquante pages du livre Stupeur et tremblements d’Amélie Nothomb. À l’aide du script PHP dont l’enregistrement se trouve sur le CD ci-joint, nous avons extrait les phrases contenant la préposition « à » du premier quart de cette œuvre. Aux pages suivantes, nous essayerons de classer toutes les occurences de cette préposition dans le groupe respectif.

Le texte étudié relève des belles-lettres, on peut donc supposer que la préposition « à » y sera utilisée le plus souvent pour construire le complément d’objet indirect. Au cas où il s’agisse de l’œuvre descriptive par exemple, le rapport de la représentation des diverses emplois de cette préposition dans le texte pourrait être changé.

Conclusion

Le centre de ce mémoire de licence consiste en description d’outils électroniques et leur exploitation dans la linguistique de corpus. Le travail s’oriente aux lecteurs qui ne disposent pas de solides connaissances sur cette problématique. Tout d’abord, il s’efforce donc d’esquisser le linguistique de corpus à grand traits, il décrit son histoire et les examples de corpus tchèques et françaises. Après il essaie d’expliquer certains termes et à la fin, dans le cas du script PHP, de montrer leur fonctionnement à l’aide d’un exemple.

La dernière partie mentionnée est focalisée sur la position de la préposition « à » dans la phrase – dans la phrase française, suivie de la phrase tchèque. Les différents exemples n’ont pas été choisis au hasard, mais ils viennent d’un certain partie d’une œuvre littéraire concrète et c’est pourquoi quelques « catégories » de l’emploi de cette préposition manquent leurs exemples (par ex. pour marquer le rapport). Le fin de ce travail se consacre à l’illustration de l’analyse faites. Les graphiques, même s’ils ne partent que d’un petit extrait du livre, montrent que la préposition « à » joue le plus souvent le rôle du complément d’objet indirect. En égard à son fonction syntaxique, on peut supposer que ce rapport serait observé aussi dans les extraits plus vastes. En abrégé, on peut dire que l’emploi de la préposition « à » dans la phrase française est si variée qu’elle devient irremplaçable pour cette langue.

En ce qui concerne des difficultés apparaissant lors de l’élaboration de ce travail, nous pouvons mentionner entre autres les problèmes avec la traduction de certains termes, viennent pour la plupart de l’informatique. L’une des raisons est sans doute le fait que les Français cherchent à créer leurs propres équivalents des termes techniques en anglais, la langue prédominante dans ce domaine.

La linguistique de corpus représente le domaine évaluant sans cesse dont les possibilités s’élargissent la main dans la main avec le progrès des technologies informatiques. Il dépend donc principalement des linguistes, des informaticiens et d’autres savants où menera la recherche dans cette discipline de recherche.

Bibliographie

Analyse et Traitement Informatique de la Langue Française. <http://www.atilf.fr/>.

Base Textuelle Frantext. <http://www.frantext.fr/>.

Bauer, L. Manual of information to accompany The Wellington Corpus of Written New Zealand English. <http://khnt.hit.uib.no/icame/manuals/wellman/INDEX.HTM>.

British National Corpus. <http://www.natcorp.ox.ac.uk/>.

C-Oral Rom Project. <http://lablita.dit.unifi.it/coralrom/index.html>.

Corpus du français parlé. <http://sites.univ-provence.fr/delic/corpus/index.html>.

CRFP. Description. <http://sites.univ-provence.fr/delic/corpus/index.html>.

Český národní korpus. <http://ucnk.ff.cuni.cz/>.

Dendien, J. – Lecomte, J. La Base Frantext. Documentation de l’utilisateur.

 <http://www.atilf.fr/atilf/produits/Tutoriel\_Frantext.pdf>. .

Dictionnaire latin-français de Félix Gaffiot. <http://www.dicfro.org/index.php>.

ELDA. <http://www.elda.org/>.

ELRA. Catalogue of Language Resources. <http://catalog.elra.info/>.

Études littéraires. <http://www.etudes-litteraires.com/>.

European Language Resources Association. Catalogue of Language Resources. <http://catalog.elra.info/>.

Evaluations and Language resources Distribution Agency. <http://www.elda.org/>.

Evert, S. The CQP Query Language Tutorial. Corpus.

< http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/>.

Francis, W. N., Kucera, H. Brown Corpus Manual.

 <http://khnt.aksis.uib.no/icame/manuals/brown/>.

International Corpus of English. <http://www.ucl.ac.uk/english-usage/ice/index.htm>.

Liste des abréviations utilisées

ACE – Australian Corpus of English

ATILF - Analyse et Traitement Informatique de la Langue Française

BMK – Brněnský mluvený korpus

BNC – British National Corpus

CERN - Conseil Européen pour la Recherche Nucléaire

CGI - Common Gateway Interface

CNRS – Centre National de la Recherche Scientifique

COBUILT - Collins Birmingham University International Language Database

C-ORAL-ROM - Corpus oral de référence intégrés pour les langues romanes

CQP – Corpus Query Processor

CRFP - Corpus de Référence du Français Parlé

CWB - Corpus Workbench

ČNK – Český Národní Korpus

DELIC - Description Linguistique Informatisée sur Corpus

DHTML - Dynamic HTML

ELDA – Evaluations and Language resources Distribution Agency

ELRA - European Language Ressorces Association

FI - Form Interpreter

FSČ – Frekvenční Slovník Češtiny

HTML - HyperText Markup Language

HTTP - HyperText Transfer Protocol

IBM - International Business Machines Corporation

ICE – International Corpus of English

IMS – Institut für Maschinelle Sprachverarbeitung

INALF - Institut National de la Langue Française

INRIA - Institut National de Recherche en Informatique et en Automatique

KDE - K Desktop Environment

KSK - Korpus soukromé korespondence

KWIC – Key Word In Context

LANDISCO – Langue Discours Cognition

LIF - Laboratoire d’Informatique Fondamentale de Marseille

MLCC - Multilingual corpora for cooperation (MLCC)

NCSA - National Center for Supercomputer Applications

OCR - Optical Character Recognition

Annexe

- Amélie Nothomb – Stupeur et tremblements, Strach a chvění (les premiers cent pages alignées de l’œuvre d’Amélie Nothomb Stupeur et tremblements et sa traduction tchèque)

- La préposition « à » - les phrases contenant cette préposition dans le document cité plus haut

- Le script PHP – l’enregistrement du script PHP créé par Mgr. Petr Vurm

- Le graphique – la clé des deux graphiques