

Pré-rapport rédigé par le
Professeur Sylviane Cardey-Greenfield
sur la thèse de Ivan Šmilauer

Intitulée « Acquisition du tchèque par les francophones :
analyse automatique des erreurs de déclinaison »

Le mémoire présenté est composé de 6 chapitres suivi de 100 pages d'annexes, soit un total de 347 pages.

L'objet de l'étude est la déclinaison du tchèque et les erreurs commises par les apprenants en vue d'alimenter une plateforme réalisée par le candidat pour l'apprentissage et la correction de la déclinaison en tchèque.

Le premier chapitre nous informe sur l'utilité de la recherche et pose le cadre méthodologique. Il fixe les objectifs de la recherche qui consiste à analyser les erreurs commises par les apprenants francophones dans des exercices de déclinaison du tchèque. L'analyse devant se faire de façon automatique, il a fallu travailler non seulement sur les formes erronées mais également sur les formes justes.

Le deuxième chapitre intitulé annotation morphologique indique comment les étiquettes et le lexique vont être choisis. Le lexique sera choisi en fonction des mots les plus fréquents de la langue. On aurait pu ici peut-être se référer aux travaux sur le français fondamental et le BASIC English. En fait ce qui paraît confus au début sur la façon de choisir le lexique s'éclaire petit à petit, on trouve la vraie réponse plus loin, le candidat va utiliser les critères discriminatoires concernant la déclinaison pour établir son corpus. Ce qui est très bien. Un bref état de l'art sur la segmentation morphologique est donné ; suite à celui-ci le candidat décrit son propre modèle car les précédents sont peu appropriés d'après le candidat pour simuler l'activité langagière (nous ne voyons personnellement pas le rapport entre activité langagière et choix de la segmentation « automatique » mais ceci pourrait être expliqué lors de la soutenance). Le problème est que tout auteur d'exercices va devoir étiqueter manuellement lui-même les textes qu'il veut utiliser.

Au chapitre trois, afin de mettre en place le jeu des étiquettes, une analyse morphologique très fine est présentée. L'analyse par type permet de dire finalement que le vocabulaire, défini d'après des critères communicationnels et pragmatiques pour les niveaux A1 et 2, reflète l'usage général de la déclinaison du tchèque. La division en type est excellente mais le lecteur ne comprend pas « critère communicationnels et pragmatiques ». En effet le candidat utilise des critères formels (comment pourrait-il faire autrement ?), bien évidemment chacun d'eux représente « un sens » dans sa définition la plus large, mais ce que la machine utilisera est bien un critère de reconnaissance par la forme. On trouve, par exemple (pages 103 et 104) la description du processus sur un exemple. Il aurait été souhaitable d'avoir un organigramme récapitulatif tous les cas, bien que nous trouvions des algorithmes appelés tests dans les pages suivantes au cas par cas. On aurait aussi aimé trouver une conclusion ou un récapitulatif à la fin de chaque chapitre.

Le chapitre quatre présente le module de diagnostic automatique qui va servir à identifier les types d'erreurs commises par les apprenants. Il est possible de générer les 3 types d'erreurs qu'un apprenant pourrait faire. Ces 3 types sont subdivisés en 9 types et leurs combinaisons (page126). Il n'y aurait que 4 types d'erreurs en réalité et le candidat les classe en 4 groupes (local, vertical, horizontal interne, horizontal externe (page127)). Une fois l'erreur trouvée et classée, elle doit être traduite en langue naturelle pour l'apprenant. La méthodologie est très bien mais, représentée de façon claire, elle n'en aurait été que meilleure.

Le chapitre cinq intitulé « évaluation » fait une critique générale du système, sont marqués les points positifs mais l'auteur n'hésite pas à donner aussi les points à améliorer (suite à 2 enquêtes), ce qui est d'une grande honnêteté scientifique.

Le chapitre six expose le fonctionnement de la plateforme destinée aux auteurs d'exercices et aux apprenants.

On trouve la conclusion au chapitre sept.

Les travaux couvrent la chaîne complète du traitement automatique qui va de la problématique à la solution implémentée.

Les quelques remarques mineures faites n'entachent en rien la grande qualité de la thèse et nous n'hésitons aucunement à donner un avis favorable pour sa soutenance.

Besançon le 9 novembre 2008

Professeure Sylviane Cardey-Greenfield
Institut universitaire de France
Université de Franche-Comté