Application d'identification de la langue d'écriture des phrases d'un texte

L'application présentée a été développée au cours du stage en ingénierie linguistique réalisé au Centre de Recherche Public - Gabriel Lippmann (Luxembourg), du 14 juin au 20 août 2004.

L'application développée permet l'identification de la langue d'écriture des phrases d'un texte. Les langues prises en compte par le prototype sont le français, l'anglais, l'allemand, la langue luxembourgeoise et le néerlandais.
Ce programme comprend deux algorithmes principaux : le premier découpe le texte en phrases et le second détermine la langue d'écriture de celles-ci. La technique utilisée pour la reconnaissance de la langue est celle des *n*-grammes.

Les différentes étapes du stage seront présentées. Les premiers jours ont été consacrés à la création de corpus qui ont servi à l'enrichissement des bases de trigrammes.
La première ébauche consistait en un système de reconnaissance par trigrammes interdits dans la langue. Ce principe a ensuite été enrichi par une technique de reconnaissance par trigrammes fréquents et une technique d'identification par mots-clés. Les résultats obtenus n'étaient cependant pas suffisants.
La version définitive du programme fait uniquement appel aux probabilités des trigrammes dans les différentes langues. Cette technique offre des résultats beaucoup plus intéressants.

L'algorithme de découpage du texte en phrases scinde le texte à chaque fois qu'il rencontre un séparateur de phrases. Le cas particulier du point nécessite certains tests qui permettent d'établir si celui-ci est utilisé comme séparateur de phrases, comme séparateur de chiffres, dans un sigle, …

Des améliorations pourraient encore être apportées (augmentation du nombre de langues prises en compte, des performances de reconnaissance ou de la technique de découpage). Cependant, le temps imparti pour le stage ne permettait pas de développer une application sans failles.

Laurent Pierret