Un protocole d'expérimentation sur les propriétés graphémiques avec l'algorithme SOM

Otman Manad*, Nourredine Aliane*, Gilles Bernard*

*Laboratoire d'Informatique Avancée de Saint-Denis (LIASD),
Université PARIS 8, 2 Rue de la Libertée, Saint-Denis, France
{manad, nourredine}@ai.univ-paris8.fr, gilles.bernard@iedparis8.net

Résumé. [1]Nous présentons une recherche sur la distribution et la classification non-supervisée des graphèmes. Nous visons à réduire l'écart entre les résultats de recherches récentes qui montrent la capacité des algorithmes d'apprentissage et de classification non-supervisée pour détecter les propriétés de phonèmes, et les possibilités actuelles de la représentation textuelle d'Unicode. Nos procédures doivent assurer la reproductibilité des expériences et garantir que l'information recherchée n'est pas implicitement présente dans le pré-traitement des données. Notre approche est capable de catégoriser correctement de potentiels graphèmes, ce qui montre que les propriétés phonologiques sont présentes dans les données textuelles, et peuvent être automatiquement extraites à partir des données textuelles brutes en Unicode, sans avoir besoin de les traduire en représentations phonologiques.

1 Introduction

Cet article présente un protocole d'expérimentation pour découvrir les propriétés orales et scripturales des graphèmes sans aucune connaissance préalable, à partir d'un corpus et à l'aide d'un réseau neuromimétique, les cartes auto-organisatrices (SOM - Self Organizing Map) (Kohonen, 1995). Un graphème est peut être définit comme la plus petite unité qui compose un document textuel, et qui correspond à un seul son en oral (phonème). Un graphème peut être composé de ponctuations ou d'un ou plusieurs caractères, par exemple : « a, c, ch, qu, eaux ».

Nos résultats montrent que les expérimentations sur les propriétés structurelles et distributionnelles des caractères peuvent donner des moyens d'accéder aux propriétés des graphèmes et aux phonèmes sous-jacents ; nous contribuons ainsi à réduire cet écart. Nous avons choisi SOM parmi les algorithmes non-supervisés, pour sa capacité à cartographier et visualiser les résultats en deux dimensions, avec une détection facile des phénomènes analysés.

2 Problématique

Parmi les digrammes, 'ca' est plus fréquent que 'tc' en raison des propriétés des consonnes et des voyelles ; mais le 'ch' est plus fréquent que le 'ct' pour une raison complètement dif-

1. Cet article présente une version remaniée et abrégée de l'article paru dans (Bernard et al., 2015)