

Réalisation d'un système de repérage automatique d'unités textuelles

Objectifs :

Le but du stage consiste à concevoir et programmer un système de repérage automatique d'unités textuelles spécifiques, dans les textes numérisés. Les textes (de langues et de domaines différents) sont fournis. Ils seront préalablement annotés d'informations de nature linguistique (catégories morpho-syntaxiques, lemmes, informations sémantiques). Les outils d'annotation¹ de texte sont fournis et peuvent être conjugués.

Pour repérer les occurrences des unités qui l'intéressent, l'utilisateur s'appuie sur des structures et indices de surface de ces unités et les formalisent au moyen de filtres ou patrons(patterns) qui sont généralement des automates ou expressions régulières.

Le système sert par exemple à repérer des noms de personne en s'appuyant sur l'observation suivante : les mots en majuscules précédés de "Monsieur", "Madame", etc. sont des noms de personnes. Entités nommées, gloses², et de façon générale tout phénomène linguistique repérable par une forme spécifique, pourront ainsi être pointés.

De tels repérages sont les "briques" des systèmes de recherche d'information dans les textes numérisés.

Spécifications :

L'entrée du système sera une description formelle de l'entité recherchée. La sortie marquera les occurrences repérées par exemple en couleur contrastée, dans les textes. Le système pourra être développé en Java, Perl/Tk ou autre. Il sera coopératif, c'est-à-dire interactif avec l'utilisateur. Par exemple, si le système marque une unité textuelle trop courte ou trop longue, l'utilisateur devra pouvoir la modifier. Autrement dit, le système est semi-automatique, il constitue une aide au repérage et il est capable d'intégrer la correction "manuelle". Ces corrections permettent d'améliorer la qualité des résultats. Elles peuvent également être analysées pour raffiner les patrons et améliorer la performance du système.

1 Exemple d'annotation réalisée par le logiciel Cordial :

N°mot	mot	lemme	Typegram	Codegram
=====	DEBUT DE PHRASE	=====		
1	il	il	38 Pp3msn	
2	y	y	38 Pp3..d	
3	a	avoir	103 Vmip3s	
4	les	le	16 Da-.p-d	
5	vrais	vrai	01 Afpmp	
6	mensonges	mensonge	25 Ncmp	
7	et	et	20 Cc	
8	les	le	16 Da-.p-d	
9	fausses	faux	03 Afpfp	
10	confidences	confidences	27 Ncfp	
11	.	.	209 Yps	
=====	FIN DE PHRASE	=====		

2 Cf. Sujet MR2 <http://www.lirmm.fr/~ducour/M2R/Stages/mela_roche_M2R.pdf>