À la croisée des langues

Annotation et fouille de corpus plurilingues

Pascal Vaillant* et Isabelle Léglise**

*Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMRS 1142),
74 rue Marcel Cachin, 93017, Bobigny cedex, France
INSERM, U1142, LIMICS, 75006, Paris, France
Sorbonne Universités, UPMC Univ Paris 06, UMRS 1142, LIMICS, 75006, Paris, France
vaillant@univ-paris13.fr
**CNRS, Structure et Dynamique des Langues (SeDyL), (UMR 8202),
7 rue Guy Môquet, 94800, Villejuif, France
leglise@vjf.cnrs.fr

Résumé. Un programme de recherche en cours sur l'étude des phénomènes de contact de langues et de leur rôle dans le changement linguistique s'attache à recueillir des corpus plurilingues, témoignant d'une grande variété de phénomènes de contact sur un échantillon suffisamment varié de langues génétiquement et typologiquement distinctes. Cet effort a impliqué le développement d'une chaîne de traitement des corpus numériques qui tienne compte des spécificités des corpus plurilingues, pour la représentation des données linguistiques, leur stockage, leur annotation, leur visualisation, et les traitements de recherche d'information. Les normes existantes ont dû être étendues pour prendre en compte l'appartenance potentielle d'unités à plusieurs langues dans les pratiques langagières plurilingues. Dans cet article, nous décrivons la manière dont a été définie la structure de ces corpus plurilingues, et la conception technique de l'unité linguistique multilingue qui préside à la fouille de données dans ces corpus.

1 Introduction

Le contact de langues est l'une des forces motrices du changement linguistique. Cette assertion, évidente lorsque l'on pense au yiddish ou aux langues créoles, est également un postulat bien connu des historiens de la langue qui ont étudié, par exemple, le passage du latin aux langues romanes, ou l'émergence de l'anglais moderne. À l'origine de ces changements, il y a nécessairement l'interaction entre des individus aux répertoires linguistiques plurilingues (Gumperz, 1982) qui, en alternant et mélangeant les langues produisent toutes sortes de variations dans l'une ou l'autre des langues et des pratiques langagières plurilingues décrites dans la littérature comme *codeswitching*, *code-mixing* et *fused lects* (Auer, 1999), *polylanguaging* (Jørgensen et al., 2011), pratiques langagières hétérogènes (Léglise, 2012). Cette multitude d'actions individuelles (Matras, 2009) prend place dans des situations sociales multilingues dans lesquelles ces variations et innovations se propagent pour progressivement mener au changement (Léglise et Chamoreau, 2013). Ainsi, les situations de multilinguisme, impliquant des