TALN 2010, Montréal, 19–23 juillet 2010

Des cartes dialectologiques numérisées pour le TALN

Yves Scherrer
Université de Genève
Rue de Candolle 5, 1211 Genève 4, Suisse
yves.scherrer@unige.ch

Résumé. Cette démonstration présente une interface web pour des données numérisées de l'atlas linguistique de la Suisse allemande. Nous présentons d'abord l'intégration des données brutes et des données interpolées de l'atlas dans une interface basée sur Google Maps. Ensuite, nous montrons des prototypes de systèmes de traduction automatique et d'identification de dialectes qui s'appuient sur ces données dialectologiques numérisées.

Abstract. This demonstration presents a web interface for digitized data of the linguistic atlas of German-speaking Switzerland. First, we present the integration of raw and interpolated atlas data with an interface based on Google Maps. Then, we show prototypes of machine translation and dialect identification systems which rely on the digitized dialectological data.

Mots-clés : Dialectologie, atlas linguistique, traduction automatique, identification de dialectes.

Keywords: Dialectology, linguistic atlas, machine translation, dialect identification.

1 Introduction : Atlas dialectologiques et traitement du langage

Depuis la fin du 19ème siècle, des études dialectologiques ont été entreprises dans de nombreuses régions linguistiques. Dans la plupart des cas, ces recherches ont donné lieu à des atlas linguistiques qui présentent les différences dialectales sous formes de cartes. Mis à part les projets les plus récents, ces ouvrages ont été réalisés sans l'apport technique de l'ordinateur, avec des cartes dessinées à la main. Cependant, de nombreuses études dialectologiques récentes (dialectométrie) font appel à un traitement de données quantitatif par ordinateur, avec l'objectif d'inférer des distances dialectales et des classifications de dialectes (Nerbonne & Heeringa, 2001; Séguy, 1973). L'utilisation de ressources dialectologiques pour des recherches quantitatives passe donc souvent par une étape de numérisation. [1] Dans la première partie de cette démonstration (Sections 2 et 3), nous discutons les enjeux de la numérisation partielle de l'atlas linguistique de la Suisse allemande et présentons ces données sur un site web dynamique.

Dans le champ du TALN, on constate également un intérêt accru pour les dialectes, avec des travaux sur l'identification de dialectes (Biadsy *et al.*, 2009), sur l'analyse syntaxique de dialectes (Chiang *et al.*, 2006; Vaillant, 2008) et sur la traduction automatique (Scherrer, 2009). Toutefois, dans la plupart de ces travaux, les dialectes sont conçus comme des entités scalaires et bien distinctes. En réalité, cette hypothèse

1. Pour une vue d'ensemble de données dialectologiques disponibles en format numérique, voir `http://www.ericwheeler.ca/atlaslist`.

est rarement valide, et on a affaire le plus souvent à un continuum dialectal avec des zones plus stables et des zones de transition. L'intégration de données géographiques permet donc de mieux tenir compte de ces réalités. Dans la deuxième partie de la démonstration, nous proposons donc un système de traduction automatique vers les dialectes suisse allemands (section 4) et un système d'identification de dialectes (section 5) basés sur les cartes dialectologiques numérisées. Ces systèmes sont accessibles sur internet.

2 Numérisation de l'atlas linguistique de la Suisse allemande

Le *Sprachatlas der deutschen Schweiz* (SDS) (Hotzenköcherle *et al.*, 1962 1997) est constitué de 8 volumes publiés entre 1962 et 1997 ; la collection des données s'est effectuée entre 1939 et 1959. Les deux premiers volumes couvrent les principales différences phonétiques et phonologiques. Le troisième volume est dédié à la morphologique flexionnelle et dérivationnelle, tandis que les cinq volumes restants couvrent le lexique dialectal. Au total, le SDS contient plus de 1500 cartes dessinées à la main, présentant des données récoltées dans 573 localités de la partie germanophone de la Suisse (et dans quelques ilôts germanophones du Piémont italien).

Le projet de numérisation présenté ici ne couvre qu'une partie de ce matériel. La sélection du matériel a été guidée par l'objectif initial de nos travaux de recherche (voir sections 4 et 5). Il s'est notamment avéré que les cartes lexicales du SDS concernent en grande partie la vie rurale des années 1940 et tendent à devenir obsolètes. À l'heure actuelle, nous avons numérisé une cinquantaine de cartes phonétiques (15% des cartes phonétiques du SDS), une centaine de cartes morphologiques (37%), et une trentaine de cartes lexicales (3%). Afin d'accélerer le processus, la complexité des cartes a été réduite de deux manières. Premièrement, les variantes dialectales apparaissant dans moins de 10 points d'enquête ont été écartées. Deuxièmement, certaines variantes phonétiques ont été regroupées puisque les conventions orthographiques que nous utilisons ne permettaient pas de les distinguer.

La numérisation proprement dite est effectuée à l'aide du logiciel ArcMap, un système d'informations géographiques. En prenant comme modèle une carte originale scannée ou photographiée (les cartes du SDS dépassent le format A3), un fichier est créé pour chacune des variantes dialectales y présente. Ce fichier contient les coordonnées des points d'enquête auxquels cette variante est recensée.

Dans la démonstration, nous montrerons comment ces données peuvent être consultées de manière interactive sur une page web. Un script Python lit les fichiers ArcMap et affiche les variantes dialectales sous forme de marqueurs sur une carte Google Maps. Toutes les démonstrations présentées ici sont accessibles à l'adresse `http://latlcui.unige.ch/~yves/`.

3 Interpolation des cartes numérisées

Les cartes originales du SDS ainsi que les cartes numérisées décrites ci-dessus représentent les phénomènes linguistiques sous forme de points ; chaque point représente une localité d'enquête. Si cette représentation en points est plus fidèle aux données initiales, une représentation en surfaces est plus proche de la réalité. Par exemple, si trois points d'enquête adjacents présentent la même forme dialectale, on peut admettre que cette forme est valable également dans les villages et hameaux qui se trouvent à l'intérieur de ce triangle.

Le passage d'une représentation en points vers une représentation en surfaces est appelé interpolation. L'objectif de l'interpolation est d'estimer la probabilité de chaque variante linguistique à chaque endroit de la Suisse allemande en fonction des informations collectées aux points d'enquête. Par exemple, si une observation A apparaît au milieu d'une zone uniforme d'observations B, on peut en déduire que A est une observation aberrante avec une probabilité basse. En revanche, même dans les endroits où B a été observé, il existe une probabilité infime que la variante A y apparaisse. En somme, l'interpolation permet de généraliser les données obtenues dans un nombre restreint de points d'enquête, par un nombre restreint d'informateurs (en général, un seul informateur par point d'enquête).

Concrètement, nous adaptons la méthode de Rumpf *et al.* (2009). Ils estiment un champ d'intensité continu pour chaque variante dialectale. Ainsi, à chaque pixel de la carte résultante, l'intensité d'une variante indique la vraisemblance que cette variante est utilisée au pixel donné. Le champ d'intensité est calculé à l'aide d'une estimation par noyau. Cette procédure est implantée à l'aide d'outils existants dans le logiciel ArcMap.

La démonstration permettra aux spectateurs de voir ces données interpolées sous forme d'images semi-transparentes, superposables sur une carte Google Maps. Cette application est accessible par l'adresse donnée ci-dessus.

4 Traduction automatique

Les données dialectologiques décrites ci-dessus sont intégrées dans un système de traduction automatique basé sur des règles. Ce système traduit un texte en allemand standard vers un dialecte suisse allemand choisi sur une carte (Scherrer, 2009).

La traduction commence par une analyse syntaxique et morphologique du texte source. Chaque mot est lemmatisé (analyse de mots composés comprise) et annoté avec des traits morphologiques et syntaxiques. Ensuite, le texte annoté est traduit mot par mot.[2] Partant de la forme de base d'un mot allemand standard, des règles phonétiques et lexicales sont utilisées pour créer une nouvelle forme de base dialectale. Un générateur morphologique dialectal crée ensuite les formes fléchies correctes. La plupart des règles phonétiques et lexicales, ainsi que les règles du générateur morphologique, sont liées à des cartes dialectologiques. Elles sélectionnent différentes variantes dialectales selon les coordonnées du dialecte cible.

Dans la démonstration, nous montrerons une page web permettant à l'utilisateur de sélectionner un dialecte cible sur une carte et de traduire un texte allemand dans ce dialecte.

5 Identification de dialecte

L'identification de dialecte, ou plus généralement l'identification de langue, est habituellement basée sur la distribution de caractères ou de n-grammes de caractères. Cette approche peut être problématique pour les dialectes dont les inventaires de phonèmes et de graphèmes sont très similaires, ou pour lesquelles des

2. Des règles syntaxiques permettant des accords à longue distance et des changements d'ordre de mots seront intégrées prochainement. Ces règles seront basées sur les données de l'atlas syntaxique du suisse allemand, actuellement en cours de rédaction (Bucheli & Glaser, 2002).

corpus d'entraînement ne sont pas disponibles.

Nous proposons une alternative basée sur une approche « sac de mots » : nous essayons d'identifier des mots dans un texte et de déterminer dans quelles régions ces formes apparaissent. Pour ce faire, nous générons d'abord une liste de mots dialectaux à partir d'une liste de mots-occurrences allemands standards, à l'aide du système de traduction automatique. En faisant cela, nous gardons la trace des cartes dialectales impliquées dans la traduction, ce qui nous permet de localiser géographiquement chacun de ces mots.

Ensuite, les mots du texte à identifier sont recherchés dans la base de données, et les cartes de chacun des mots sont combinés. En fin de compte, chaque phrase est représentée par une carte qui visualise la distribution de probabilités de la phrase dans les différentes régions suisses. Nous montrerons à l'aide de quelques exemples comment différents dialectes suisse allemands peuvent être identifiés grâce à cet outil en ligne.

Remerciements

Une partie de ces travaux a été réalisée durant mon séjour à l'Université Columbia de New York, financé par le Fonds National Suisse de la Recherche Scientifique (bourse No. PBGEP1-125929). Je tiens à remercier Owen Rambow pour les nombreuses discussions stimulantes au sujet de mes recherches.

Références

BIADSY F., HIRSCHBERG J. & HABASH N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athènes.

BUCHELI C. & GLASER E. (2002). The Syntactic Atlas of Swiss German Dialects : empirical and methodological problems. In S. BARBIERS, L. CORNIPS & S. VAN DER KLEIJ, Eds., *Syntactic Microvariation*, volume II. Amsterdam : Meertens Institute Electronic Publications in Linguistics.

CHIANG D., DIAB M., HABASH N., RAMBOW O. & SHAREEF S. (2006). Parsing Arabic dialects. In *EACL'06 : Proceedings of the Eleventh Conference of the European Chapter of the Association for Compuational Linguistics*, p. 369–376, Trento.

HOTZENKÖCHERLE R., SCHLÄPFER R., TRÜB R. & ZINSLI P., Eds. (1962-1997). *Sprachatlas der deutschen Schweiz.* Bern : Francke.

NERBONNE J. & HEERINGA W. (2001). Computational comparison and classification of dialects. In W. VIERECK & H. GWOSDEK, Eds., *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*, volume 9, p. 69–83. Alessandria : Edizioni dell'Orso.

RUMPF J., PICKL S., ELSPASS S., KÖNIG W. & SCHMIDT V. (2009). Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik*, **76**(3).

SCHERRER Y. (2009). Un système de traduction automatique paramétré par des atlas dialectologiques. In *Actes de TALN 2009*, Senlis.

SÉGUY J. (1973). La dialectométrie dans l'atlas linguistique de la Gascogne. *Revue de linguistique romane*, **37**, 1–24.

VAILLANT P. (2008). Grammaires factorisées pour des dialectes apparentés. In *Actes de TALN'08*, p. 159–168, Avignon.