Séminaire La Bretagne Linguistique

Vendredi 2 mars 2012

Faculté des lettres Victor-Segalen, Brest
Salle des thèses (C219, 2e étage)

9h30 : Accueil des participants

10h00 : Guylaine BRUN-TRIGAUD, « Le "THESAURUS OCCITAN" : une base de données multimédia dédiée aux dialectes occitans »

11h00 : Elisabetta CARPITELLI, « L'*Atlas Linguistique Roman* : enjeux et perspectives »

12h00 : Pause déjeuner

14h00 : Elisabetta CARPITELLI et Daniel LE BRIS, « Concordances linguistiques entre aires atlantique et romane »

15h00 : Adrien DESSEIGNE et Pierre-Yves KERSULEC, « Les enquêtes dialectologiques de la Société d'Ethnolinguistique Bretonne : Eléments de présentation du Questionnaire Grammatical 1500 »

UBO université de bretagne occidentale

CRBC EA 4451

Le "THESAURUS OCCITAN" : une base de données multimédia dédiée aux dialectes occitans

Guylaine BRUN-TRIGAUD
CNRS / Laboratoire Bases-Corpus-Langage,
Université de Nice Sophia-Antipolis

Le *Thesaurus Occitan* ou THESOC est une base de données multimédias, qui contient notamment :

- des données linguistiques et péri-linguistiques issues d'enquêtes de terrain : cartes et carnets d'enquêtes des Atlas linguistiques, monographies, enregistrements sonores, documents iconographiques ;
- des données linguistiques procédant d'analyses déjà réalisées : lemmatisations, morphologie, étymologie, microtoponymie ;
- des données bibliographiques ;
- des outils d'analyse : représentations cartographiques, instruments d'analyse diachronique, procédures de cartographie comparative, instruments d'analyse morphologique.
- un Module Morpho-Syntaxique (MMS), détaillé plus loin ci-dessous.

Centralisé à Nice dans le cadre du laboratoire UMR 6039 du CNRS « Bases, Corpus, Langage », sous la direction de Jean-Philippe Dalbera, il s'agit d'un programme interuniversitaire qui associe différentes équipes.

Les données brutes figurant dans le THESOC répondent à deux critères qui conditionnent leur implémentation dans la base :

- les faits doivent être précisément localisés, ce qui constitue une condition essentielle pour l'étude de la variation diatopique. Cela permet notamment par la suite de générer dynamiquement et automatiquement des cartes linguistiques sur demande.
- les faits doivent provenir de sources orales. En effet, la philosophie est ici d'intégrer des faits recueillis sous forme orale (avec transcription en API), ce qui garantit la réalité des faits considérés. De plus, le THESOC permet de faire entendre les sons enregistrés au cours des enquêtes, ce qui permet à l'utilisateur de la base de contrôler la transcription proposée.

Le THESOC est donc un objet à géométrie variable qui envisage toutes sortes d'exploitations grâce à des menus spécifiques, et qui intègre toutes sortes de documents, si bien que le THESOC se présente comme un outil offrant à la fois (mais toujours séparément) des données linguistiques quasi brutes, des données ayant fait l'objet d'analyses et de traitements et des outils d'investigation. L'intérêt d'un tel outil réside également dans le fait qu'il peut évoluer en permanence selon les besoins des utilisateurs.

Dans le cadre de cette présentation, certains aspects lexicaux communs entre les parlers occitans et les parlers bretons seront particulièrement mis en évidence.

English Version

The *Thesaurus Occitan* (abbreviated THESOC) is a multimedia database, which contains, among other things:

- linguistic and linguistic-related data from field works : maps and survey notebooks from the *Atlas linguistiques*[1], monographies, audio records, pictures ;
- linguistic data coming from former analyses : lemmatisation, morphology, etymology, microtoponymy ;
- bibliographical references ;
- tools for linguistic analyses: maps generator, instruments for diachronic analyses, comparative cartography procedures, morphological analysis instruments.
- a Morpho-Syntax Module (MMS), detailed further below.

Centralised in Nice within the laboratory UMR 6039 « Bases, Corpus, Langage » (attached to the CNRS), this inter-university program associates different teams, upon the direction of Pr. Jean-Philippe Dalbera.

For some raw data to be included in the THESOC, it must match two criteria:

- linguistic facts must be precisely located. This constitutes an essential condition for diatopic variation studies. In particular, this condition later allows to dynamically and automatically generate linguistic maps on demand.
- linguistic facts must come from oral sources. Indeed, our philosophy here is to integrate linguistic facts collected under oral form (with IPA transcription), which guarantees the reality of these facts under consideration. Moreover, the THESOC allows hearing of the audio tracks recorded during the field works, thus giving to the user the possibility to control or to check the proposed transcription.

After a presentation of the THESOC's lexical database, we will focus more precisely on the Morpho-Syntax Module, especially designed for syntactic and morpho-syntactic analysis of Occitan dialects. This module contains both oral texts (including ethnotexts) and single sentences, such as answers to morphosyntactic questionnaire. The localisation of these texts and sentences enables on the long term a comparison between different dialects on a syntactical basis, thus opening new perspectives for dialectology.

Even if this module was originally conceived for oral texts processing, its tagger and parser are still able to process written texts so far as they are written in a familiar or popular style, close to oral register. By the way, we will explain how we made these computer tools manage linguistic variation in all its dimensions (dialectal variation, graphical variation, inflectional variation, etc.).

One can say the THESOC is a variable geometry database which considers all kinds of exploitations thanks to specific menus, which integrates all kinds of documents. Therefore, the THESOC looks like a tool offering both (but always separately) almost raw linguistic data and data resulting from former analysis and treatments, plus investigation tools. The advantage of such a database lies also in the fact that it can evolve and be updated permanently to satisfy users' needs.

---

2. *Atlas Linguistiques de la France par régions*, éditions du C.N.R.S.

L'*Atlas Linguistique Roman* : enjeux et perspectives

Elisabetta CARPITELLI
UMR 6039 Bases, Corpus, Langages, ISHS de Nice
Université de Nice-Sophia Antipolis

L'*Atlas Linguistique Roman* (ALiR) — tout comme l'*Atlas Linguarum Europae* (ALE) — avant d'être un ouvrage de géolinguistique naît comme un grand chantier de collaboration internationale entre spécialistes de géolinguistiques qui mènent ensemble, grâce à une collaboration constante, une réflexion sur la reconstruction des systèmes linguistiques de l'aire romane. Une attention particulière est réservée depuis le début à la reconstruction lexicale, à la lumière des études sur la motivation lexico-sémantique. Le but de la communauté scientifique n'est pas de publier des données "brutes" à une échelle plus vaste – dans ce cas, l'échelle d'un domaine linguistique – par rapport à celle de la majorité des atlas, mais de cartographier et commenter l'interprétation des données dialectales figurant sur les atlas nationaux et régionaux déjà publiés. L'ALE a été le premier atlas à proposer de manière systématique et théoriquement explicite des cartes motivationnelles qui, depuis le début de l'entreprise, se sont révélées particulièrement intéressantes en relation au domaine de la zoonymie dialectale. L'accent a été mis sur l'identité ou les ressemblances des représentations culturelles et idéologiques entre espaces linguistiques, même très éloignés du point de vue géographique et génétique, plutôt que sur les différences formelles, de surface. Ce type d'analyse nécessite des compétences multiples : celles des spécialistes de chaque domaine linguistique ainsi que de celles qui proviennent de l'apport de l'anthropologie, de l'ethnologie, de l'histoire des religions, de l'archéologie ; ainsi les cartes des deux atlas sont souvent le fruit d'une véritable recherche interdisciplinaire. L'ALiR, né comme filiation de l'ALE, a opté pour la publication de volumes thématiques de cartes et de commentaires motivationnels, à partir d'une sélection de questions d'atlas qui ne répète pas celle de l'atlas continental. Si au départ, l'homogénéité du domaine, du point de vue génétique — par rapport à l'hétérogénéité de l'espace recouvert par l'ALE — avait laissé craindre aux chercheurs un résultat escompté, au contraire, la comparaison s'est révélée fructueuse : grâce à la grande quantité de données comparées et cartographiées simultanément, pour la première fois, il a été même possible de reprendre en considération une série d'étymologies non élucidées.

Dans le cadre de cette rencontre, on présentera le chantier, avec une réflexion globale qui n'ignore pas les problèmes de fonctionnement qui accompagnent le travail des chercheurs, en essayant de monter l'importance que la dialectologie et la géolinguistique revêtent encore dans le domaine de la reconstruction.

Les enquêtes dialectologiques de la Société d'Ethnolinguistique Bretonne : Eléments de présentation du Questionnaire Grammatical 1500

Pierre-Yves KERSULEC
Université Rennes 2

Adrien DESSEIGNE
Université Paris Descartes

La Société d'Ethnolinguistique Bretonne, créée en 2010, s'est donné comme objectif d'effectuer un travail de documentation et de mise en valeur des dialectes bretons. A cet effet, un questionnaire grammatical, comportant quelque 1500 phrases à traduire, a été élaboré par les membres de l'association afin de recueillir auprès des locuteurs de breton les caractéristiques morphosyntaxiques de leur parler.

L'objectif qui est le nôtre dans cette communication est de présenter les tenants et les aboutissants de ce questionnaire, à la lumière de nombreuses enquêtes accomplies sur le terrain, et au regard de la publication à venir d'un projet en cours d'aboutissement (questionnaire complet effectué à Berné).

Nous évoquerons d'une part la nature du questionnaire en elle-même, ainsi que les conditions de déroulement des enquêtes nécessaires à sa réalisation. Par ailleurs, nous définirons les enjeux spécifiques de ce projet, ainsi que ses limites ou écueils éventuels.

Nous conclurons, enfin, notre communication, en abordant la question des perspectives d'exploitation des données recueillies jusqu'à présent.

Concordances linguistiques entre aires atlantique et romane

Elisabetta CARPITELLI
Université de Nice

Daniel LE BRIS
Université de Brest

Nous avons effectué plusieurs comparaisons entre les zones celtiques et gallo-romanes en Europe occidentale. En effet, le croisement et/ou la complémentarité dans certains cas des domaines roman et celtique permet d'éclaircir des formes lexicales devenues opaques dans la première aire linguistique mais toujours transparentes dans la seconde, et vice-versa. Il met à jour l'existence de continuateurs, d'aboutissants restés jusqu'alors inaperçus. Ce travail collectif basé sur une analyse motivationnelle du lexique met en évidence la présence de concordances aréales en zone atlantique. Il contribue d'une certaine manière à porter un autre regard sur la répartition dialectale dans cette partie de l'Europe et propose de véritables perspectives de recherche prometteuses.