Présentation du programme Euroslav2010

Evangelia Adamou, Walter Breu

PRÉSENTATION DU PROGRAMME *EUROSLAV 2010*
Base de données électronique de variétés slaves menacées dans des pays européens non slavophones

1. Présentation générale et objectifs

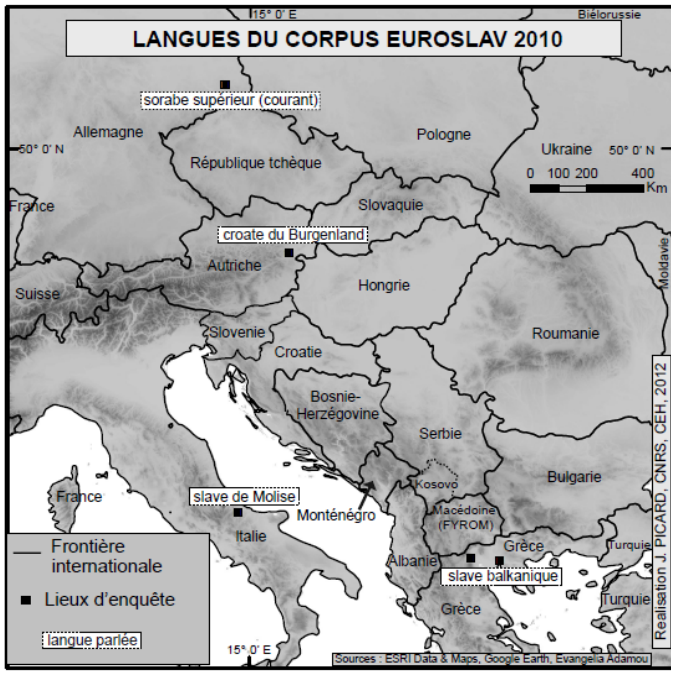
EuroSlav 2010 : *Base de données électronique de variétés slaves menacées dans des pays européens non slavophones* est un programme de recherche franco-allemand[1]. Il a permis la création d'une base de données électronique valorisée totalisant plus de 5 heures de corpus oral synchronisé avec les transcriptions, les annotations et les traductions libres. Le corpus EuroSlav 2010 porte sur des variétés slaves en voie de disparition parlées dans quatre pays européens : il s'agit notamment du sorabe supérieur courant, en Allemagne, du croate de Burgenland, en Autriche, de variétés slaves balkaniques, en Grèce (nashta et Hrisa), et du slave de Molise, en Italie (na-našu) : voir carte 1[2].

Les variétés slaves du Sud et de l'Ouest intégrées dans ce projet sont parlées dans des Etats européens dont la langue officielle n'appartient pas à la famille slave (grec, italien, allemand). Dans la plupart de cas, les locuteurs abandonnent leur variété slave au profit des langues de l'État, plus utiles pour leur ascension sociale et faisant partie de leur identité nationale. Les situations varient tout de même en fonction des pays. En Grèce par exemple, la question des populations slavophones reste tabou, alors qu'en Autriche, en Allemagne et en Italie la présence de locuteurs slavophones est reconnue, et qu'un enseignement leur est proposé dans le cadre scolaire. Toutefois, cet enseignement est souvent facultatif et peu suivi (notamment en Italie). Dans tous les cas, les formes standardisées et puristes sont celles qui sont promues ou enseignées, formes qui ne sont pas représentatives des variétés parlées traditionnellement dans ces localités. Les données proposées dans le cadre du projet EuroSlav 2010 sont précieuses à deux titres : d'une part ces microlangues risquent de disparaître dans les 30 ans à venir ; d'autre part, dans certains cas, comme en Grèce, elles sont pratiquement inconnues et non documentées.

---

[1] Financement Agence Nationale de la Recherche (ANR-09-FASHS-025) & Deutsche Forschungsgemeinschaft (DFG : BR 1228/4-1), 2010-2012. Coordinateur du projet pour la partie allemande : Walter Breu (Universität Konstanz). Coordinateur du projet pour la partie française : Evangelia Adamou (CNRS). Équipe allemande : Lenka Scholze (Universität Konstanz), Mia Barbara Mader Skender, Jasmin Meinzer, Maria Utschitel (doctorantes), Stuart Cunningham (traduction anglaise), Marie-Antoinette Confignal (traduction française). Equipe française : Georges Drettas (chercheur, CNRS), Séverine Guillaume (ingénieure d'études en informatique, CNRS), Yordanka Kozareva (post-doctorante), Gwenaël Le Bras (assistant), Margaret Dunham (traduction).

[2] Walter Breu a fourni les données de vernaculaires slaves parlés en Italie du Sud, enregistrées dans trois localités : Acquaviva Collecroce, Montemitro et San Felice del Molise. Lenka Scholze a fourni les données du sorabe supérieur courant, parlé en Allemagne et, avec Maria Utschitel, celles du croate du Burgenland, parlé en Autriche. Georges Drettas a contribué par les enregistrements qu'il avait réalisés dans les années 1970 à Hrisa en Grèce. Enfin, Evangelia Adamou a fourni les données pour le vernaculaire parlé à Liti (nashta).



Carte 1. Les localités représentées dans le projet EuroSlav 2010

La plupart de projets de documentation n'incluent pas les variétés en voie de disparition appartenant à des familles de langues de grande diffusion, par exemple les variétés slaves dont il a été question dans le programme ANR-DFG EuroSlav 2010. Les projets de dialectologie slave ou européenne qui seraient plus susceptibles d'inclure ces données ont des objectifs différents, bien que complémentaires, de ceux du programme EuroSlav 2010. En effet, les études dialectologiques ont une forte tradition, fondée notamment sur l'étude de la phonétique et du lexique, et ne sont pas encore bien développées en ce qui concerne la morphologie et la syntaxe (on doit mentionner ici des projets qui ont modernisé la dialectologie en y intégrant aussi la syntaxe, comme c'est le cas du projet dirigé par Sjef Barbiers *Syntactic Atlas of the Dutch Dialects* (*SAND*), ou bien la base de données *Romani Morphosyntax* dirigée par Yaron Matras et Victor Elšik). Par ailleurs, les textes élaborés dans une perspective dialectologique sont rarement glosés ou analysés selon les normes de la linguistique générale, et s'adressent surtout aux spécialistes de ces langues. Dans le cadre du programme EuroSlav 2010, il est désormais possible pour les linguistes généralistes et les typologues de consulter un corpus finement annoté de vernaculaires slaves riche en phénomènes de contact de langues, pouvant alimenter les recherches menées dans ce domaine.

2. Méthodologie et aspects techniques

Du point de vue méthodologique, la collecte de données de vernaculaires en voie de disparition vise souvent une authenticité perdue : par exemple, pour la description des variétés parlées en Grèce, la collecte de données a souvent été faite auprès d'immigrés et de réfugiés politiques ayant vécu pendant de nombreuses années en République de Macédoine ou en Bulgarie, en faisant abstraction de l'influence des langues slaves standard apprises entretemps par les locuteurs ; pour le slave de Molise, le croate de Burgenland et le sorabe supérieur courant

(totalement différent du sorabe supérieur littéraire, à tous les niveaux linguistiques) il y a souvent des influences puristes (de la part d'acteurs locaux et étrangers) dans les rares textes publiés. Cette approche « puriste » sera sans doute réfutée, pour une part grâce aux textes sonores et aux analyses du programme EuroSlav 2010.

Une des originalités du programme EuroSlav 2010 a donc été son approche théorique linguistique et typologique appliquée aux vernaculaires slaves. Sur le plan méthodologique, les enregistrements ont été effectués *in situ*, et les données ont été analysées en prenant en compte leurs conditions de production réelles. Au niveau de l'analyse, les emplois contemporains et la variation observée chez un ou plusieurs locuteurs de la communauté constituent un objet d'étude. En effet, ces variétés ne sont pas de simples dialectes des langues les plus proches, pour lesquels il suffirait de signaler les écarts comme des traits archaïques (croate, bulgare, macédonien) : elles présentent de nombreuses innovations, voire des phénomènes inattendus pour des langues slaves, et, surtout, elles sont riches en phénomènes dus au contact de langues : cf. par exemple Breu (2008) pour la création de systèmes d'article dans le slave de Molise et le sorabe supérieur courant, ou Breu (2005) pour les changements dans le système aspectuel de ces langues, ou bien encore Scholze (2008) pour une vue d'ensemble des « singularités » du sorabe supérieur courant.

Une grande partie des enregistrements avaient déjà été numérisés avant le début du projet ; c'est le cas des données de la variété de Hrisa des années 1970. Ainsi une phase de numérisation n'a pas été nécessaire pour ces textes dans le cadre du projet EuroSlav 2010. En général, les fichiers audio archivés dans le cadre du programme sont au format WAV pour une qualité optimale, avec quelques exceptions. Des versions compressées (en MP3 monophonique) sont proposées lors de la consultation sur internet pour une meilleure fluidité.

Pour les transcriptions, on peut distinguer trois cas de figure : 1. transcription inexistante ; 2. un texte transcrit sous forme manuscrite (travaux antérieurs au projet, sur des données recueillies lors de missions anciennes) ; 3. une transcription saisie sous une forme informatisée. Nous avons adopté une transcription phonétique en Alphabet Phonétique International, avec un codage Unicode pris en compte notamment par XML. Les variétés d'Italie, d'Allemagne et d'Autriche, qui sont par ailleurs dans une certaine mesure enseignées dans le cadre d'un enseignement bilingue, ont aussi été notées selon les conventions orthographiques des différents standards, afin qu'elles soient lues par les populations concernées, qui ont l'habitude de lire dans ces alphabets. L'homogénéisation des gloses a posé quelques difficultés : des cadres théoriques différents, des systèmes linguistiques variés avec des besoins d'annotation différents, ont compliqué cette tâche. Toutefois, les participants se sont mis d'accord pour employer les règles élaborés à Leipzig par B. Bickel, B. Comrie et M. Haspelmath (Leipzig Glossing Rules), qu'ils ont enrichies en fonction des besoins des langues étudiées. Ce choix de suivre les normes de la linguistique générale a été fait afin de rendre les données accessibles pour des études de typologie et de contact de langues. En effet, les documents archivés sont des outils de recherche et non seulement des objets de patrimoine. Ils se doivent donc d'être suffisamment complets afin de pouvoir servir comme base pour le maximum de types de recherche, comme par exemple les travaux sur la prosodie.

Pour la saisie nous avons utilisé le logiciel Interlinear Text Editor (ITE, M. Jacobson), un éditeur de textes structurés en XML qui permet de gloser un document en présentant les phrases sous une forme interlinéaire, et en entretenant un lexique de toutes les gloses utilisées. La version 1.0 de XML a été finalisée par le W3C (World Wide Web Consortium) en février 1998, ce qui constitue une garantie de très grande diffusion et de développement rapide des outils de gestion (interrogation, édition, etc.).

Les chercheurs du programme EuroSlav 2010 ont été confrontés à plusieurs problèmes lors des transcriptions, des gloses et des traductions des textes. D'une part, le logiciel ITE ne permettait pas d'intégrer certaines fonctionnalités qu'ils souhaitaient utiliser. Par exemple, le marquage des emprunts et du code-switching n'étaient pas prévus dans ITE. Une autre difficulté concernait les traductions dans des langues multiples qu'ITE ne pouvait pas prendre en charge de manière efficace. Enfin, il est vite apparu nécessaire aux participants de pouvoir consulter les fichiers XML avant leur mise en ligne. Des solutions techniques ont été proposées pour toutes ces questions par S. Guillaume, permettant au programme de poursuivre ses objectifs, mais une actualisation des outils de saisie est indispensable. Des « passerelles » ont été mises en place pour que les fichiers XML au format Lacito puissent être repris sur le logiciel Elan pour les chercheurs qui souhaitent les développer davantage.

La synchronisation *(time alignment)* de la transcription et du son au niveau de la phrase a été faite avec le logiciel SoundIndex (développé par M. Jacobson pour le Lacito au CNRS). SoundIndex, un logiciel connecté à ITE, associe un éditeur de son et un éditeur de texte XML.

Une traduction a été faite en anglais, en français et en grec pour les corpus des vernaculaires parlés en Grèce. Une traduction en allemand, français et anglais a été faite pour le corpus des variétés parlées en Autriche et en Allemagne, et une traduction en italien, allemand et anglais pour celles parlées en Italie. Nous avons pensé que la traduction dans les langues majoritaires des pays dans lesquels ces vernaculaires sont parlés était utile non seulement pour la lecture des textes mais aussi parce qu'elle mettait plus clairement en parallèle les deux langues en contact et soulignait la présence des emprunts (notés en italique dans la transcription phonétique du corpus et marqués par des astérisques dans la langue majoritaire en question). Enfin, toutes les gloses ont été faites en anglais pour permettre une meilleure comparaison des corpus entre eux. Pour une description plus détaillée de la segmentation morphologique, la présentation phonétique (segmentale et prosodique) et la structure générale du corpus EuroSlav 2010 cf. Breu, Adamou (2011).

Des métadonnées ont été renseignées pour chaque fichier audio et chaque fichier XML : le lieu, la date, la nature de l'enregistrement, etc. et le langage de balisage (partie informatique). Ces méta-données sont doublement encodées en DCMI (Dublin Core Metadata Initiative) et en OLAC (Open Language Archives Community), format qui précise l'interprétation de certaines étiquettes DCMI pour le domaine des archives de parole. Toutes ces informations sont accessibles en utilisant un protocole définit par l'OAI (Open Archives Initiative).

Le corpus EuroSlav 2010 est intégré au programme *Pangloss* du Lacito-CNRS, une archive publique contenant 1230 enregistrements en 71 langues, dont 450 documents annotés (http://lacito.vjf.cnrs.fr/archivage/presentation.htm). Les don-

nées de l'archive sont structurées selon les normes actuelles de l'informatique, dans un format ouvert. Les fichiers sont archivés de manière pérenne dans le cadre d'une infrastructure numérique d'accès aux données et documents des sciences humaines et sociales mise en place par le CNRS, le Très Grand Équipement ADONIS (http://www.tge-adonis.fr/). De cette façon, la pérennité de l'accès à la base de données est assurée. Nous avons également la garantie que le corpus EuroSlav 2010 suivra les progrès technologiques.

L'utilisateur accède au document par l'intermédiaire d'un « navigateur » standard (Firefox, Explorer, etc.) ; il peut écouter le son correspondant à un segment choisi de la transcription, ou encore écouter tout l'enregistrement pendant que la transcription défile sur l'écran. Il peut choisir d'afficher ou non les traductions dans différentes langues, le mot-à-mot aligné avec la transcription, etc.

L'accès au son est assuré soit par un *player*, soit par un applet Java, soit par un plug-in. Le player choisi établit une communication avec le navigateur via un script Java. Ce dernier redirige les requêtes de l'utilisateur vers le player (arrêt du son, démarrage du son à telle ou telle phrase, etc.). C'est lui aussi qui va demander au navigateur de mettre en valeur un segment particulier en réaction aux messages (activation, inactivation) envoyés par l'applet.

La diffusion des données est assurée par une architecture qui s'appuie sur les technologies du web. Côté serveur, une machine héberge sur ses disques toutes les données archivées (fichiers XML, WAV). Pour diffuser ces données, un serveur web (Apache) a été installé sur cette machine, c'est lui qui va assurer la communication entre les machines clientes et la machine serveur. Hormis le serveur web et des données d'archives, un processeur de styles (Xalan) ainsi qu'un certain nombre de feuilles de styles (XSL) ont été ajoutés afin de pouvoir traduire à la volée les documents d'archives en documents XHTML directement interprétables par les machines clientes.

Côté utilisateur le seul outil nécessaire est un navigateur web, plus les ressources utiles pour rendre correctement les caractères Unicode des textes (polices), ainsi que celles pour rendre correctement le son des enregistrements (le Java Media Framework ou un plug-in audio).

Le programme EuroSlav 2010 rend accessibles en ligne les enregistrements sonores valorisés en proposant plusieurs modes de consultation :

(1) Il est possible de suivre graphiquement phrase par phrase la transcription du texte sonore.
(2) Il est possible d'afficher le mot-à-mot, constitué de gloses morphosyntaxiques permettant de découvrir la grammaire de la langue.
(3) Il est également possible de choisir de une à trois langues de traduction accompagnant le texte.
(4) Il est aussi aisé de repérer les passages énoncés dans la langue de contact (nommés alternances codiques ou code-switching) ainsi que les mots qui proviennent de la langue de contact actuelle ou non (indiqués en italique).

Ci-dessous, un extrait du corpus de Liti tel qu'il apparaîtra sur la plateforme Pangloss du Lacito :

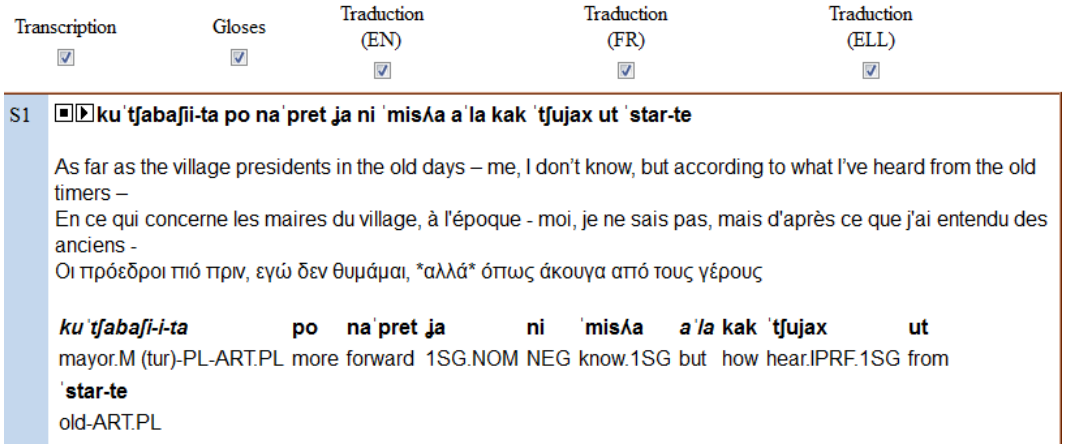


Figure 1. Extrait du corpus du nashta tel qu'il s'affiche sur la plateforme Pangloss

3. Résultats dans le domaine du contact de langues

L'approche adoptée dans ce programme de recherche est une approche empirico-déductive (ascendante, ou plus couramment appelée en anglais *bottom-up*), et qui constitue un grand enjeu pour la linguistique du 21e siècle. En effet, la linguistique théorique et formelle se convertit aussi à l'étude des grands corpus (Siemund 2011). Par ailleurs, la linguistique de laboratoire cherche à développer le travail à partir de corpus aussi proches que possible de la communication ordinaire et d'un contexte naturel *(ecologically valid)*. Il s'agira donc, dans un deuxième temps, d'exploiter et d'analyser les corpus oraux déjà constitués dans le cadre du programme EuroSlav 2010 afin de contribuer aux études de contact de langues.

Parmi les objectifs scientifiques qui découlent de l'exploitation des corpus constitués, nous allons, entre autres, tester l'hypothèse typologique proposée par Yaron Matras selon laquelle les éléments empruntés sont les mêmes à travers les langues du monde, car déterminés par des critères cognitifs (Matras 2007 ; Matras 2009). Suivant ce principe, Matras établit différentes hiérarchies d'empruntabilité à partir de la fréquence avec laquelle une catégorie est affectée par le changement induit par le contact. Il établit ainsi des hiérarchies implicationelles qui se confirment dans plusieurs langues.

Nous essaierons, dans le cadre d'une publication commune en préparation, de comparer les emprunts et les convergences des systèmes en contact en fonction des caractéristiques typologiques des langues slaves et non-slaves concernées (ex. disposant d'un article défini, d'une distinction aspectuelle, etc.) mais aussi en fonction de l'intensité de contact. On note que la langue source est non seulement la variété standard mais le plus souvent la variété locale. Ce travail sera appliqué sur les corpus constitués dans le programme EuroSlav 2010. Quelques résultats préliminaires découlant de la comparaison des corpus sont présentés ci-dessous.

Par exemple Matras (2007 : 54) observe que, dans plusieurs langues du monde, les connecteurs seront empruntés dans l'ordre suivant :

but > *or* > *and*

Le tableau suivant illustre les résultats dans le corpus EuroSlav 2010 avec des exemples confirmant cette hiérarchie (en nashta et na-našu), et d'autres qui l'infirment (en sorabe SC, en croate BL et à Hrisa). On note + les éléments issus

de la langue de contact, – la présence d'un connecteur slave, et (–) un élément mixte.

	mais >	ou >	et
sorabe SC	–	–/+	–
na-našu	+	+/(–)[3]	+/–
croate BL	–	–/+	–
nashta	+	+	–
Hrisa	–	–	+

Tableau 1. Les connecteurs dans les corpus EuroSlav 2010

En regardant de plus près, on observe qu'en nashta, l'adversatif du grec, *ala*, a remplacé l'adversatif du turc, *ama,* qui était emprunté dans pratiquement toutes les variétés slaves balkaniques. En revanche, *ama* est employé avec les deux valeurs du grec, avec une valeur conditionnelle et une temporelle. Dans Adamou & Vanhove 2006 étaient étudiés les deux contours intonatifs de ces emplois, un contour montant pour le conditionnel et un descendant pour le temporel. Il est intéressant de noter que les locuteurs du nashta n'ont pas emprunté le coordonnant du grec *ke*, alors qu'il a été emprunté à Hrisa, d'après le corpus de Drettas des années 70. Il est intéressant de noter qu'à Hrisa en revanche, l'adversatif du turc *ama* n'a pas été remplacé par celui du grec. Cette diversité pose des questions quant à la date d'introduction des emprunts dans les vernaculaires slaves, questions que nous essaierons d'approfondir dans l'ouvrage prévu à l'issue du programme EuroSlav 2010.

Le cas du na-našu est d'un intérêt particulier pour la hiérarchie d'empruntabilité proposée pour les particules de réponse par Matras (2007). Selon cette hiérarchie, il y aurait une tendance dans les langues du monde à emprunter d'abord la particule de réponse positive, puis la particule de réponse négative. En na-našu, *keja* 'oui', qui est la forme la plus fréquente, est compris par les locuteurs comme une forme slave (en opposition à *si* emprunté à l'italien, *se* emprunté à la variété locale). Toutefois, *keja* ne se retrouve dans aucune autre langue slave (en Dalmatie non plus), et il pourrait s'agir d'une forme anciennement empruntée aux dialectes italiens des Abruzzes[4]. Les autres langues du corpus EuroSlav 2010 confirment cette hiérarchie, puisqu'en sorabe supérieur courant et à Hrisa aucune particule de réponse n'est empruntée, alors qu'en nashta c'est la particule de réponse positive qui est empruntée (voir le tableau ci-dessous).

[3] En slave de Molise (na-našu) la situation du connecteur 'ou' est complexe : d'un côté on trouve l'emprunt italien *o* (marqué par +), mais de l'autre on a la forme *ol,* attesté aussi en Dalmatie, dans laquelle *o* semble avoir fusionné avec la forme slave *(i)li.*

[4] On trouve des formes comme *chéjja* 'celle (chose)', par exemple à Teramo (Giammarco 1968 : 517), qui pourrait être la source d'un emprunt signifiant 'oui'. Par ailleurs il y a aussi d'autres caractéristiques du slave de Molise qui permettent de soutenir un passage par les Abruzzes au cours de l'immigration des slaves en Molise au XVI^{e} siècle.

	oui >	non
sorabe SC	–	–
na-našu	–/+	+/–
croate BL	–/+	–
nashta	+	–
Hrisa	–	–

Tableau 2. Les particules de réponse positive et négative dans les corpus EuroSlav2010

En nashta, la particule phatique empruntée au grec, *ne* 'oui', coexiste avec la négation homophone du nashta (qui, elle, est maintenue), *ne* 'non' (réalisée aussi [ɲa]). Dans un corpus enrichi et synchronisé avec le son, il est possible de mener des études au niveau de la prosodie. Bien que nous n'ayons pas mené d'étude sur cet aspect, il est toutefois clair que la particule de réponse positive et négative en nashta ont deux contours intonatifs distincts. Le corpus EuroSlav 2010 permet de valider cette observation, puisque non seulement aucune occurrence du slave *da* n'y apparaît, mais que la particule phatique *ne* intervient dans des textes qui sont pratiquement sans alternance codique avec le grec.

Une autre hiérarchie d'empruntabilité concerne les numéraux. Greenberg (1978 : 290) faisait déjà observer que 1 est toujours retenu, et que 2 et 3 sont plus souvent retenus que les numéraux plus élevés. Enfin, il notait : "If an atomic numeral expression is borrowed from one language into another, all higher atomic expressions are borrowed." (Greenberg 1978 : 289). Matras (2007 : 51, 52) propose quant à lui les hiérarchies suivantes :

> Numéraux élevés 1000, 100 > supérieurs à 20 > supérieurs à 10 > supérieurs à 5 > inférieurs à 5
> Numéraux dans des contextes formels > numéraux dans des contextes informels

Bien que les corpus EuroSlav 2010 ne contiennent pas les occurrences de tous les numéraux, ils constituent une base sur laquelle on a pu faire une analyse plus détaillée des langues à l'étude. Le tableau ci-dessous distingue les numéraux de 1 à 4, de 5 à 10, ceux qui sont supérieurs à 10, et dégage 100 comme un cas à part. Les données du na-našu confirment l'intérêt de séparer les nombres entre 5 et 10, qui présentent une variation entre formes slaves et emprunts, des nombres entre 11 et 99, qui s'utilisent seulement comme emprunts[5]. En croate du Burgenland on observe que certains locuteurs emploient exclusivement les formes slaves et d'autres varient entre les formes slaves et les formes allemandes, notamment pour les numéros supérieurs à 10.

[5] La distribution des nombres en slave de Molise est très compliquée, dans la mesure où elle ne dépend pas seulement de la génération du locuteur mais aussi de l'objet compté. En outre il ne s'agit pas seulement d'un choix entre deux formes, mais trois, par ex. *pet* = *čing* = *čingue* 'cinq', car à côté des nombres slaves hérités et des nombres empruntés anciennement aux dialectes locaux de la Molise, on trouve aussi les nombres empruntés plus récemment à l'italien, qui, eux aussi, participent aux règles de distribution ; pour une recherche détaillée sur les nombres cf. Breu (en préparation).

	1-4	5-10	> 10	100
sorabe SC	–	–	–	–
na-našu	–	+/–	+	+/–
croate BL	–	–	–/+	–/+
nashta	–	+/–	+/–	–
Hrisa	–	–	–	–

Tableau 3. Les numéraux dans les corpus EuroSlav 2010

Un autre aspect important du contact de langues est la convergence des structures (morphologie, ordre des constituants, sémantique lexicale, compatibilités syntaxiques). Nous allons étudier ces convergences en suivant l'approche de Breu (2008) selon laquelle il y aura convergence entre deux structures si la même catégorie fonctionnelle existe déjà dans les deux langues et fait partie d'une structure polysémique dans la langue dominante L_2 (voir tableau 4).

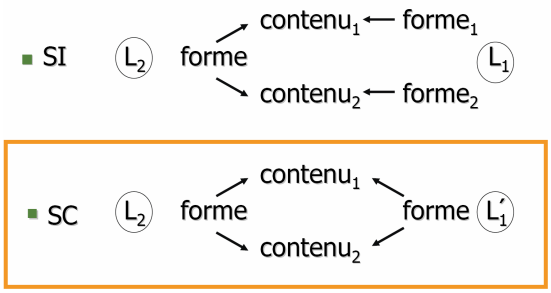


Tableau 4. L'adaptation de la structure sémantique des langues en contact

La convergence entre les langues en contact peut se faire à travers une perte de catégorie grammaticale dans les langues minoritaires, mais dans de nombreux cas on assiste au développement d'une nouvelle catégorie dans ces langues sous l'influence de la langue de contact majoritaire. Par exemple, en ce qui concerne l'article indéfini, l'italien et l'allemand disposent d'un numéral 1 grammaticalisé aussi comme article indéfini. Sur la base du modèle italien et allemand, l'article indéfini a été grammaticalisé en slave de Molise et en sorabe SC. Le croate du Burgenland offre un exemple d'un stade intermédiaire de la grammaticalisation d'un article indéfini.

Lorsqu'un modèle polysémique n'existe pas dans la langue majoritaire, Breu soutient que cette structure ne sera pas affectée et donne l'exemple de l'article défini : d'une part le sorabe SC reproduit le modèle allemand, qui dispose d'un démonstratif et d'un article défini partageant la même forme. Le sorabe SC, sous l'influence de l'allemand, élargit dans ce cas les fonctions du démonstratif slave existant *tón*. En revanche, comme en italien les démonstratifs et l'article défini sont distincts au niveau de la forme, le slave de Molise ne développe pas d'article défini à partir du démonstratif. Cette approche, qui explique la convergence des structures morphosyntaxiques en s'appuyant sur le modèle polysémique disponible dans la langue dominante (L_2), sera étayée par les données du corpus EuroSlav 2010. Au niveau morphosyntaxique on trouve des exemples de convergence

pour le futur, le mode irréel, le résultatif, l'aspect verbal, les constructions explétives, le passif, les cas, le genre, etc.

4. Conclusion et perspectives

Un cadre de linguistique générale, tel que celui du programme EuroSlav 2010, permet aux données des vernaculaires de trouver leur juste place en leur assurant une visibilité internationale. Outre les slavisants, de nombreux linguistes s'intéressant au contact de langues et à la typologie pourront ainsi accéder à des données orales naturelles, analysées et annotées, de ces variétés slaves peu documentées et en voie de disparition. Rendre les données analysées accessibles à tous grâce aux programmes de documentation semble la meilleure manière de procéder. Ces vernaculaires vont bientôt disparaître, et il est de notre responsabilité de rendre publiques et d'archiver ces données précieuses en les préparant au mieux à une exploitation future.

Plusieurs perspectives s'ouvrent pour l'élaboration du corpus constitué dans le cadre du programme EuroSlav 2010 pour l'étude de nouveaux aspects linguistiques. Ainsi, en collaboration avec Martine Toda, ingénieure à l'IR CORPUS (2011-2012), nous avons commencé à exploiter les corpus du programme EuroSlav 2010 afin d'extraire tous les formants/spectres relatifs à certains sons, dont la réalisation phonétique est particulièrement intéressante. M. Toda a, par exemple, testé le logiciel EasyAlign (développé par Jean-Philippe Goldman, Université de Genève) avec des premiers résultats encourageants pour le corpus nashta. Il s'agit d'un alignement automatique qui permettra de faire dans un deuxième stade une analyse semi-automatique des formants, du spectre, ou des transitions formantiques, pour déterminer les caractéristiques dynamiques de ces transitions. Il est certain que le fait de s'appuyer sur les corpus naturels de langues à tradition orale participe d'un renouvellement des études linguistiques.

Bibliographie

Adamou, E. 2012. Verb morphologies in contact: evidence from the Balkan area. In: Vanhove, M., T. Stolz, H. Otsuka, A. Urdze (eds.). *Morphologies in contact*. Berlin, 143-162.

Adamou, E., M. Vanhove 2006. What does prosody tell us about syntax? Présentation orale dans *Second Conference on the Syntax of the World's Languages.*

Breu, W. 2005. Verbalaspekt und Sprachkontakt. Ein Vergleich der Systeme zweier slavischer Minderheitensprachen (SWR/MSL). In: S. Kempgen (ed.), *Slavistische Linguistik 2003, Referate des XXIX. Konstanzer Slavistischen Arbeitstreffens.* München, 37-95.

Breu, W. 2008. Развитие систем артиклей при полном контакте славянских меньшинств с немецким и итальянским языками. In: Kempgen, S., K. Gutschmidt, U. Jekutsch, L. Udolph (eds.), *Deutsche Beiträge zum 14. Internationalen Slavistenkongress Ohrid 2008.* München, 75-88.

Breu, W. (en préparation). Zahlen im totalen Sprachkontakt. Das Numeralsystem des Moliseslavischen. In: Reuther, T. (ed.), *Slavistische Linguistik 2012. Referate des XXXVIII. Konstanzer Slavistischen Arbeitstreffens*. München.

Breu, W., E. Adamou 2011. Slavische Varietäten in nichtslavophonen Ländern Europas. Das deutsch-französische Gemeinschaftsprojekt *EuroSlav 2010*. In: Kempgen, S., T. Reuther (eds.), *Slavistische Linguistik 2010* (=*Wiener Slawistischer Almanach* 37). München, 53-84.

Giammarco, E. 1968. *Dizionario Abruzzese e Molisano*. Volume Primo A–E. Roma.

Greenberg, J. 1978. Generalizations about Numeral Systems. In: Greenberg, J., C. Ferguson, E. Moravcsik (eds.), *Universals of Human Language*. Vol. 3: Word Structure. Stanford, 249-295.
Matras, Y. 2007. The borrowability of structural categories. In: Matras, Y., J. Sakel (eds.), *Grammatical borrowing in cross-linguistic survey*. Berlin, New York, 31-73.
Matras, Y. 2009. *Language contact.* Cambridge.
Scholze, L. 2008. *Das grammatische System der obersorbischen Umgangssprache im Sprachkontakt. Mit Grammatiktafeln im Anhang.* Bautzen.
Siemund, P. (ed.). 2011. *Linguistic Universals and Language Variation.* Berlin, New York.

Paris (CNRS) Evangelia Adamou
(adamou@vjf.cnrs.fr)

Konstanz Walter Breu
(Walter.Breu@uni-konstanz.de)