Pour une approche multilingue des langues

Patrice Pognan

LES NOUVEAUX CAHIERS FRANCO-POLONAIS

ASPECTS SOCIOLOGIQUES ET ANTHROPOLOGIQUES DE LA TRADUCTION

No 7/2008

Collection :
LES NOUVEAUX CAHIERS FRANCO-POLONAIS, N° 7

ASPECTS SOCIOLOGIQUES ET ANTHROPOLOGIQUES DE LA TRADUCTION

Sous la rédaction de

Zofia Mitosek
Anna Ciesielska-Ribard

CENTRE DE CIVILISATION POLONAISE (UNIVERSITE DE PARIS-SORBONNE)
FACULTE DE LETTRES POLONAISES (UNIVERSITE DE VARSOVIE)

Paris – Varsovie 2008

PATRICE POGNAN

Langues, Logiques, Informatique et Cognition (LALIC-Certal)
Université de Paris-Sorbonne et INALCO
France

POUR UNE APPROCHE MULTILINGUE DES LANGUES

Nous avions abordé lors du colloque en Sorbonne le multilinguisme essentiellement sous l'aspect de la traduction (en général) et de la terminologie multilingue dans le cadre de la traduction automatique. Nous souhaiterions insister ici sur deux points qui nous semblent mériter des développements importants :

1. l'analyse automatique des langues et ses applications d'une part à la terminologie et à la recherche d'information et, d'autre part, à l'éventuelle réalisation automatique de dictionnaires.
2. l'apprentissage non pas d'une langue, mais d'un système linguistique et de ses variations avec l'élaboration des outils adéquats tant à la présentation pédagogique qu'aux recherches linguistiques nécessaires. Nous envisageons ce type de démarche pour le groupe des langues slaves de l'Ouest (avec un éventuel regard sur d'autres langues slaves), pour les langues berbères où il est nécessaire de faire apparaître le système linguistique (coopération avec Miloud Taïfi et le réseau marocain RueLing pour le tachelhit / chleuh, le tamazight et le tarifit / rifain), pour des langues turques (turc, azéri, turkmène, kazakh, ouzbek, ouïghour).

Nous tenterons de montrer que ces deux points relèvent des mêmes stratégies et d'un même savoir linguistique.

1. Des situations conduisant au multilinguisme

1.1. *Le manque de documentation.* Comment apprendre l'azéri si l'on a besoin de s'approprier cette langue ? Il existe en Occident bien peu d'outils pour un tel apprentissage, par exemple un petit lexique anglais-azéri et azéri-anglais et un tout petit manuel de conversation précédé d'une micro-grammaire et de micro-lexiques dans les deux sens, le tout publié par Hippocrene. Une solution radicale et efficace est de se lancer dans l'apprentissage du turc fort bien documenté et ce quelle que soit la langue de l'apprenant.

Il en va de même pour l'apprentissage du « tamazight », le berbère du Maroc central, pour lequel nous sommes démunis en matière de supports pédagogiques. Nos travaux avec Miloud Taïfi, auteur du dictionnaire tamazight – français, dans le cadre d'actions intégrées franco-marocaines et d'un FSP, nous obligent à avoir une certaine connaissance de cette langue. Ici aussi le meilleur moyen consiste à passer par une des langues les plus proches, le « tachelhit » (ou chleuh), berbère du sud-ouest du Maroc pour lequel nous possédons de bonnes descriptions, des grammaires, des lexiques, des manuels d'apprentissage dont certains avec cassettes, bref de quoi apprendre. Ce choix est plus judicieux, s'il est fait seul, que celui du « taqbaylit » (kabyle) plus distant que le tachelhit.

1.2. *La nécessité scientifique.* Les techniques développées dans nos travaux d'analyse automatique, en particulier du tchèque, sont afférentes à la reconnaissance automatique de formes. Elles sont donc très liées au code écrit et à une compréhension très fine de la phonologie et de la morphologie tant diachroniques que synchroniques. Sur ces bases il est possible de mettre en évidence de nombreux éléments de ce que Jean-Marie Zemb appelait « l'autoéclairage » des mots. Mais il est des choses peu perceptibles au sein d'une langue parce que le phénomène est ténu ou bien, tout simplement, parce qu'on ne sait pas le voir. La connaissance d'autres langues, voisines ou non, peut permettre de percevoir ce que l'on ne voyait pas. Nous en donnerons deux exemples en tchèque où la lumière a été apportée par l'observation du haut-sorabe.

Des exemples précédents, il ressort la nécessaire connaissance d'une langue pour une autre : le turc pour l'azéri, le tachelhit pour le tamazight et, dans des conditions différentes, le haut-sorabe pour le tchèque. Ce type de démarche peut s'étendre et se systématiser : faire d'abord ressortir, comme nous l'avons fait abondamment en berbère avec M. Taïfi, le système linguistique qu'il faudra s'approprier en premier et s'appuyer sur une, voire deux langues, pour un apprentissage pluriel des langues du groupe. Ainsi, nous développons sur la base du tchèque une initiation aux langues slaves de l'Ouest, nous élaborons des travaux de lexicologie dans les trois parlers berbères du Maroc (tachelhit, tamazight et tarifit) et envisageons des travaux portant simultanément sur le turc, l'azéri, le turkmène, l'ouzbek, le kazakh et l'ouïghour.

1.3. *L'inadéquation pédagogique.* L'écriture chinoise a un rôle particulier : elle est le système écrit commun à toutes les langues chinoises. Ceci est très important puisque les locuteurs de langues chinoises différentes interprètent la chaîne écrite suivant leur chaîne parlée propre. D'autres langues utilisent également les caractères chinois, en premier lieu le japonais dont les kanji

(ayant deux séries de « lectures » – chinoise et japonaise) sont les caractères chinois, mais aussi les langues qui employaient autrefois le chinois, à savoir le sino coréen et le sino-vietnamien. Les étudiants de japonais et de coréen apprennent les caractères chinois. Il est nécessaire de réaliser des logiciels d'apprentissage de l'écriture chinoise qui soient communs à l'ensemble de ces langues.

L'ensemble de ces démarches va dans le sens de l'attention actuellement portée au multilinguisme. L'Union européenne, malgré des soubresauts schizophrènes, apporte un soutien réel aux langues européennes et au multilinguisme, y compris sous la forme de l'intercompréhension. En effet, à côté des formulaires, des demandes et des rapports de contrats industriels ou de recherche inexorablement, voire pour certains « naturellement » écrits en anglais (ce qui malheureusement pèse très lourd dans la balance de l'apprentissage et de la connaissance d'autres langues européennes – français, espagnol, allemand... – en tant que langues secondes ou tierces), l'Union européenne développe des programmes-cadres de grande envergure en faveur des langues, tels que les programmes Lingua, Socrates.

Sous l'égide de ces différents programmes se développent plusieurs projets de multilinguisme et d'intercompréhension. C'est Claire Blanche-Benvéniste qui a donné le coup d'envoi de ce nouveau courant de recherche avec le système Eurom4, méthode d'enseignement simultané de quatre langues romanes. Cette recherche fondatrice fut publiée en 1997 à Florence. Le tout récent congrès « Diálogos em intercompreensao » (dont les actes sont publiés au choix des auteurs en portugais, espagnol, italien, français et anglais) a réuni à Lisbonne en septembre 2007 une grande partie des acteurs et des projets de ce domaine nouveau : EU&I, Galanet, Minerva, ICE (InterCompréhension Européenne), ILTE, Eurocom, Intercom, Euromania.

2. Analyse automatique du tchèque et notion de calculabilité

La langue tchèque, langue slave de l'Ouest, est un exemple d'écriture phonologique (d'autres exemples sont le coréen – réforme du roi Sedjong au XV^{e} siècle et le turc – réforme d'Atatürk en 1928). A la différence du sorabe et du polonais, le tchèque possède une correspondance presque parfaite entre la chaîne parlée et la chaîne écrite : à un phonème correspond un graphème et à un graphème correspond un phonème. Cela est dû en premier lieu à l'élaboration de l'écriture glagolitique par Constantin au IX^{e} siècle pour les besoins de l'Empire de grande Moravie (on remarquera que la caractéristique de bi-univocité entre graphie et phonie introduite par cette écriture demeure dans les langues

slaves écrites en cyrillique), puis à la réforme introduite par Jan Hus (son traité « Ortographia Bohemica » parait en 1412, republié en 1968 à Wiesbaden) qui remplace les digraphes des diverses écritures de l'époque par des caractères diacrités. Cette solution se retrouve dans toutes les langues slaves écrites en latin, sauf le polonais qui ne l'utilise que (très) partiellement. Elle a été explicitement reprise par le lituanien (qui a légèrement changé le code). Cette écriture est souvent utilisée pour transcrire le russe en latin. Elle est également la base majoritaire de la graphie latine du berbère.

Il convient de faire remarquer que l'écriture n'est qu'un code chargé de transcrire au mieux la chaîne parlée préexistante. Même au sein d'un même système linguistique, les décisions prises au niveau de la représentation écrite peuvent avoir des conséquences diverses, en particulier sur la calculabilité de la langue. Prenons l'exemple du tchèque et du slovaque, langues sœurs qui constituent un sous-groupe / sous-système au sein des langues slaves de l'Ouest (les deux sorabes constituent un autre sous-groupe de même que le polonais et le kachoube). En tchèque, le « e mouillé » est représenté par un « e » surmonté d'un háček (« ě »). Un adjectif tchèque est un mot qui possède une désinence toujours longue (désinence constituée par une diphtongue ou possédant une voyelle longue), ce qui est le cas fondamentalement aussi en slovaque. En slovaque dont le code graphique est issu du code tchèque, le « ě » n'existe pas, car il est supposé que les suites « de », « te » et « ne » sont suffisamment caractéristiques pour exprimer de manière non ambiguë une prononciation « dě », « tě » et « ně ». D'autre part, le slovaque possède une loi dite de rythme qui empêche la succession de deux syllabes longues (la seconde est abrégée) et ceci quelle que soit la valeur de la longueur, par exemple valeur de désinence d'adjectif. Le croisement de ces phénomènes met en évidence des résultats très différents en tchèque et en slovaque. Soient les deux suites qui se correspondent en tchèque et en slovaque :

	adjectif masculin	adjectif féminin	adjectif neutre	adverbe
tchèque	krásný	krásná	krásné	krásně
slovaque	krásny	krásna	krásne	krásne
	beau	*belle*	*beau*	*de belle manière*

Alors qu'en tchèque la graphie est classificatoire, en slovaque la « calculabilité » n'est plus assurée. On se retrouve dans une situation où la machine est confrontée à une ambiguïté systémique et où les apprenants étrangers sont en difficulté : il est nécessaire de comprendre le texte pour pouvoir l'interpréter (dans un cas, adjectif avec application de la loi de rythme, dans l'autre cas,

adverbe où il y a économie du « ě » devant « n ») et pour pouvoir le lire à haute voix. On se retrouve – sous une forme infiniment plus simple – dans la situation de lecture de l'arabe non voyellé.

La troisième période est celle du « Renouveau national » où des générations de linguistes depuis Dobrovský jusqu'à Gebauer sont intervenus sur la langue, en particulier au niveau d'une mise en ordre morpho-sémantique de la suffixation. De nombreux suffixes réfèrent actuellement (avec très peu d'exceptions) à un champ sémantique déterminé :

– *« -iště », à deux exceptions près renvoie à un lieu ouvert : « hřiště » est un lieu ouvert où l'on peut « hr- » jouer = terrain de jeu*,

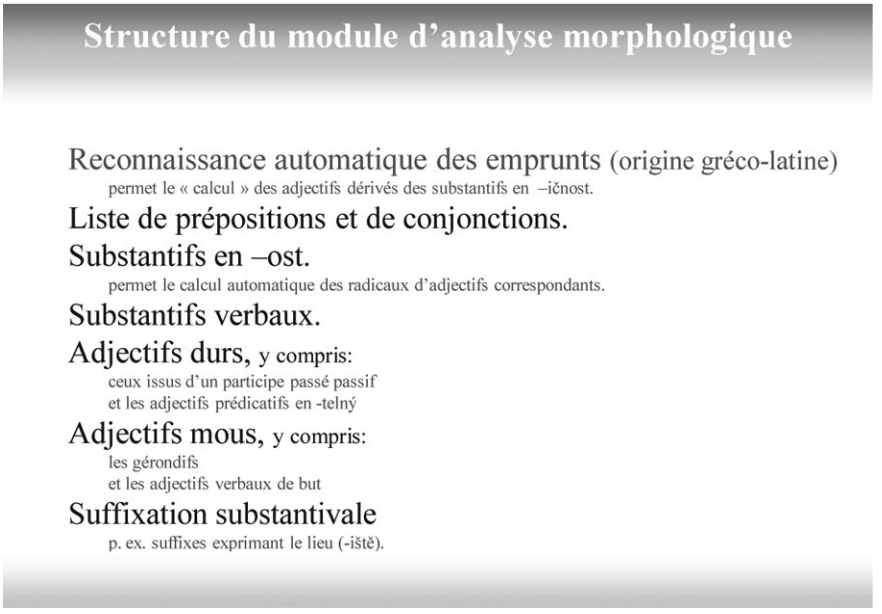
– « -ovna », à un lieu fermé : « knihovna » *la bibliothèque*, « -árna » et « -írna » à des lieux où l'on vend, fabrique quelque chose, où l'on se livre à une activité en général : « lékárna » *la pharmacie*, « tavírna » *la fonderie...*

On consultera avec profit l'énorme travail de Bares sur les noms en « -dlo ». L'analyse de tels suffixes permet non seulement une analyse morphologique, mais elle permet de plus de donner automatiquement une classification sémantique sommaire.

A ces caractéristiques très favorables à une analyse de formes, il faut rajouter une autre caractéristique primordiale qui apporte beaucoup à ce type d'analyse : c'est l'opposition de longueur entre voyelles brèves et voyelles longues. La longueur a souvent été formée par contraction, phénomène historique ayant fortement marqué la langue tchèque. L'utilisation synchronique de tels phénomènes pour l'analyse automatique du tchèque a une efficacité renforcée par la connaissance précise de la grammaire historique qui se révèle très utile pour la connaissance précise d'une langue, mais absolument nécessaire pour la compréhension d'un groupe de langues.

Cette caractéristique de longueur permet de pouvoir dire automatiquement que tout mot terminé par « -ý » (sauf deux exceptions actuelles et une autre maintenant inusitée) est un adjectif, de modèle dur au masculin singulier nominatif et peut-être accusatif si le substantif auquel se rapporte l'adjectif est inanimé. De la même manière, un mot terminé par « -ů » sera – sans que l'on connaisse la forme de nominatif singulier (consonantique ou vocalique) – un substantif masculin au génitif pluriel.

Ce ne sont ici que quelques exemples élémentaires. L'analyse s'appuie essentiellement sur la reconnaissance automatique des adjectifs (durs et mous) qui est moins complexe et donc plus accessible que l'analyse des catégories du substantif et du verbe. La structure générale du module d'analyse morphologique est la suivante :



Dans un système basé sur la reconnaissance des formes, il est important de ne pas avoir d'interférences avec des structures étrangères. C'est pourquoi le premier sous-module de l'analyse morphologique est consacré à la reconnaissance des emprunts gréco-latins qui peuvent être utilisés dans un texte de langue tchèque. Cette reconnaissance utilise deux systèmes :

– *celui des éléments appartenant exclusivement au système linguistique étranger*, par exemple les graphèmes « g » et « ó ». Ils ont tous deux des caractéristiques particulières : la grammaire historique fait apparaître que le « g » slave s'est transformé en tchèque, en slovaque et en haut-sorabe en « h », ce qui a pour conséquence que tout mot comprenant un « g » ne peut pas être tchèque (car sinon à la place du « g » il y aurait un « h »). De même, le « ó » s'est transformé en tchèque en diphtongue « uo » qui s'est par la suite transformée en « ů » (en slovaque, c'est un « ô » qui est le résultat de l'évolution du o long slave, mais toujours prononcé en tant que diphtongue « uo »). Par conséquent, tout « ó » apparaissant dans un texte tchèque marque une origine étrangère du mot le contenant. C'est pourquoi il est très regrettable que certaines réformes de l'orthographe suppriment la longueur sur le « o » de certains mots étrangers suivant que les experts en charge de la réforme prononcent ou non ce « o » de manière longue. Cette longueur a une justification historique et une utilité synchronique pour des travaux automatisés sur la langue. Elle permet aussi de montrer qu'une alternance de type « dvůr » / « do dvora » *cour* n'est en fait qu'une alternance de longueur « ó/o ».

– *la violation des lois et règles tchèques* : ainsi des mots dans lesquels intervient le non-respect de la cohérence de mouillure (une consonne ou une voyelle tchèque peut être classée en dure ou molle – concept qui semble dépassé en linguistique slave actuelle, mais qui est d'une remarquable efficacité en traitement automatique de la langue en tant que formalisation immédiate) des suites consonne – voyelle. Ainsi, le mot « historie » apparaît immédiatement comme étranger par la non-palatalisation de « h » et de « r » devant « i » (i mou) et par la présence d'une diphtongue « ie » qui en tchèque aurait été contractée en « í » (la seule diphtongue autochtone est « ou », les diphtongues « au » et « eu » sont étrangères : « auto », « eufonie »). Si ce mot avait été tchèque, il aurait abouti à une forme en « zistoří » ! De même, le mot « cykl » qui correspond à une structure qui peut être tchèque (celle d'un verbe au passé tel que « tekl » *il a coulé*) est reconnu comme étranger par la violation de la suite nécessaire consonne molle – voyelle molle (nous avons ici une consonne molle [ceci est « vrai » en tchèque : « Poláci » *les Polonais*, mais « faux » en polonais : « Polacy »] suivie d'une voyelle dure).

L'analyse automatique de la morphologie tchèque met en lumière des connaissances souvent négligées, par exemple :

1. – En premier lieu, un tel système d'analyse est sous-tendu aux niveaux phonologique et morphologique par une connaissance de la grammaire historique dans ses grands développements :

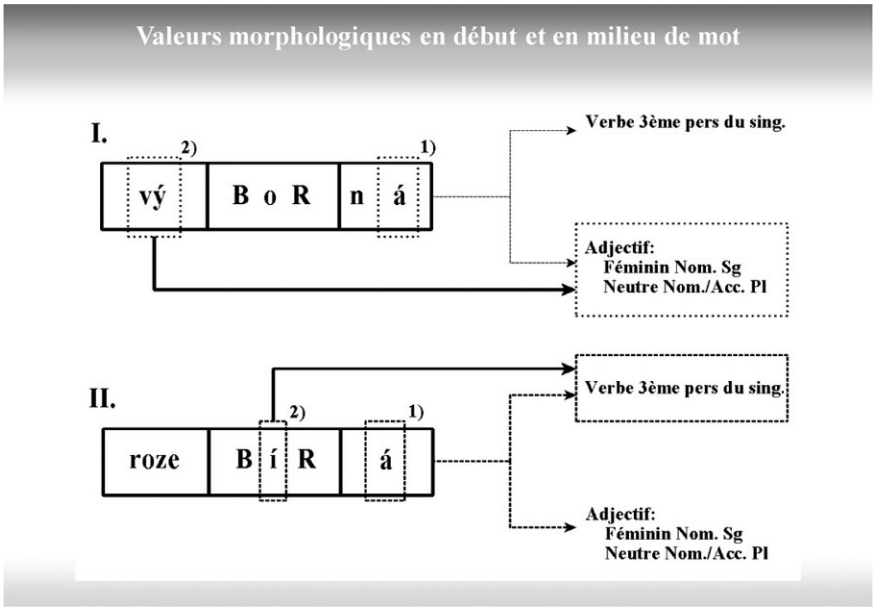
- Du protoslave à la fin du X^{e} siècle :
  - Métathèse des liquides (évolution de : ***tort***, ***tolt***, ***tert***, ***telt***)
  - Contraction
  - Disparition et vocalisation des jers
  - Evolution des nasales
- De la fin du X^{e} siècle à la fin du XIV^{e} siècle :
  - Passage g ⇒ h
  - Evolution du r mouillé en ř
  - Transformations ‘a ⇒ ě + ‘u ⇒ i
  - Dépalatalisation
- De la fin du XIV^{e} siècle à la fin du XVI^{e} siècle
  - Transformation du « u long » en diphtongue : ú ⇒ ou
  - Contraction ie ⇒ í
  - Contraction uo ⇒ ů
  - Changement aj ⇒ ej

2. – Les langues slaves, comme l'avait déjà montré Dobrovský, répugnent, à des degrés divers, à débuter un mot par une voyelle. Ainsi, le tchèque ne possède qu'une douzaine de mots-outils débutant par « a », deux mots débutant par « e » et « i », ce qui veut dire que la masse lexicale pour laquelle les mots débutent par « a », « e » et « i » est d'origine étrangère, en fait gréco-latine. Nous verrons plus loin, à l'heure de la calculabilité relative des langues slaves de l'Ouest,

que le haut-sorabe ne débute jamais un mot autochtone par une voyelle, ici aussi à quelques exceptions près.

Ainsi, un mot dans un texte en langue tchèque qui débute par « a », « e » ou « i » et n'appartient pas à la liste d'exceptions est un mot d'origine étrangère : « angažmá », « embargo », « integrovaný ». En haut-sorabe, les mots commençant par une voyelle, sauf exception, sont d'origine étrangère.

Le tchèque est une langue à flexion externe, c'est-à-dire que ce sont des suffixes désinentiels qui donnent la fonction du mot, suffixes situés donc à la fin du mot. De manière surprenante, d'autres parties du mot participent à l'analyse morphologique : la préfixation et la voyelle support de la racine, ce que nous présentons dans le schéma qui suit.



3. – Dans le premier cas, nous sommes en présence d'une désinence ambiguë « -á » qui représente soit un verbe (de 5ème classe, à la 3ème personne du présent), soit un adjectif (féminin nominatif singulier ou neutre pluriel au nominatif ou à l'accusatif). La reconnaissance de cette désinence ne suffit donc pas à déterminer la catégorie grammaticale du mot.

Par contre, il existe une série de préfixes longs qui correspondent aux formes brèves (do-, na-, při-, pro-, po-, u-, vy-, za-). Lorsque ces préfixes longs sont en tête du mot, à une dizaine d'exceptions près, le mot ne peut pas être de nature verbale. Dans le cas présenté ci-dessus, « výborná » *excellente* ne peut pas être un verbe et sera répertorié en tant qu'adjectif.

4. – Dans le second cas du schéma ci-dessus, nous avons la même ambiguïté de désinence, mais le préfixe est bref et malheureusement la relation n'est pas réversible : alors qu'un premier préfixe à voyelle longue indique de manière quasi-univoque une dérivation non verbale, le premier préfixe à voyelle brève n'a aucune valeur particulière pouvant entrer aussi bien dans une dérivation verbale que non verbale. Dans le cas présenté ici, la voyelle de la racine, le « í » – sauf deux exceptions – indique un verbe de 5ème classe, imperfectif, à la 3ème personne du singulier du présent. C'est donc cette valeur qui sera retenue. Le seul inconvénient est que pour pouvoir trouver cette valeur, il est nécessaire de conduire une analyse morphématique automatique complète.

Ces connaissances ont été acquises par l'étude de sources scientifiques concernant le tchèque, par le dépouillement de grammaires, de dictionnaires et de textes, par la coopération avec des équipes tchèques. D'autres connaissances, inattendues, ont été apportées par l'étude du haut-sorabe. Nous en citerons deux qui sont caractéristiques :

1. – Lors de l'étude de la métathèse, la comparaison entre le haut-sorabe et le tchèque fait apparaître une dissymétrie patente :

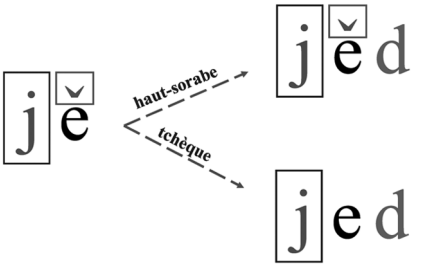
D'un archaïque « korva », on obtient « krowa » vache en polonais et « kráva » en tchèque, ce qui représente la transformation normale entre polonais et tchèque. En haut-sorabe c'est « kruwa », ce qui correspond au tchèque et au polonais. Par contre, là où nous avons en tchèque « ret » *lèvre*, nous avons en haut-sorabe « ert » *bouche* ; là où nous avons « žláza » *glande* en tchèque, nous avons « žałza » en haut-sorabe, c'est-à-dire que dans ces deux cas, la métathèse n'a pas eu lieu en haut-sorabe.

Cette remarque est importante : elle a attiré notre attention sur le fait qu'en tchèque la métathèse a été totale (c'est pour cela que les Tchèques possèdent « Labe », alors que les Français et les Allemands ont « Elbe »). Cela veut dire que toutes les formes originellement :

(consonne) {e/o} {l/r} consonne
sont devenues en tchèque :
(consonne) {l/r} {e/a} consonne.

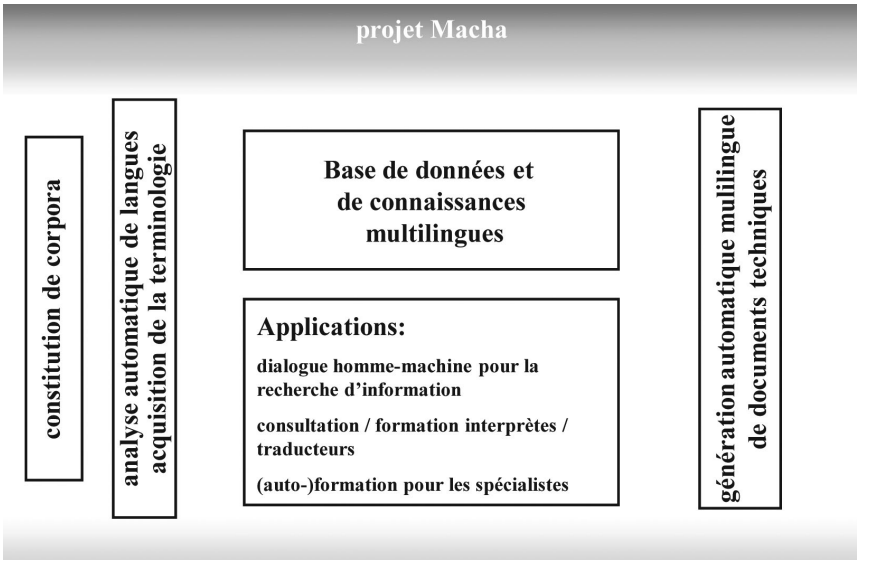
Il en résulte que les mots qui présentent en leur sein une structure du premier type sont vraisemblablement étrangers au tchèque. Nous sommes actuellement en train de faire les dépouillements de vérification correspondants, afin de compléter le module de reconnaissance des emprunts.

2. – En tchèque, pour écrire un « e mouillé », nous avons deux solutions : soit « ě » c'est-à-dire un « e » surmonté d'un háček, par exemple « běh » *la course*, soit une graphie « je » qui est obligatoire en début de mot. On écrit ainsi « jed » *poison*. En haut-sorabe, le poison s'écrit « jěd ».



A l'examen de cas semblables, il semble que le tchèque refuse la double mouillure qui est possible en haut-sorabe. Dans le jeu de la double mouillure, le háček s'efface devant le « j ». Si cette observation se révélait être fondée (ce que nous essayons de vérifier), cela aurait des conséquences intéressantes sur certaines interprétations de la morphologie.

Les techniques d'analyse automatique que nous venons de présenter ont été appliquées avec succès à d'autres langues : slovaque (Diana Lemay), anglais (Dominique Noël), italien (Sarah Labat-Jacquemin), malais (Bali Ranaivo), japonais (Nadine Rayon). Ce type de programme est destiné à faire de la recherche d'information, de la recherche de terminologie dans les textes. La mise en parallèle de tels programmes permettrait de monter un système multilingue de recherche de terminologie. Nous avons eu un tel projet, nommé Macha (machinisme agricole), qui a été – pour le moment – abandonné après la disparition de notre partenaire industriel tchèque Agmeco, issu de l'ancien Centre de recherche en machinisme agricole.



Il existe une autre application d'un tel système d'analyse réellement intéressante. Si le système est capable d'analyser sans dictionnaire toute la morphologie de la langue, il peut alors fabriquer automatiquement des dictionnaires. En effet, si le système est apte à catégoriser les mots d'un texte, au lieu de fournir les résultats sur une unité de sortie (papier, écran), il est tout à fait envisageable de les charger dans une base de données dont la teneur peut être ensuite vérifiée et complétée par le(s) concepteur(s). Un environnement de programmation tel que Python dispose d'un système de base de données SQLite et peut aisément permettre ce type de projet.

3. Apprentissage d'un système linguistique et de ses variations

Nous insisterons dans cette partie sur la nécessité de s'appuyer sur une connaissance minimale de la grammaire historique (l'idéal étant, bien sûr, une connaissance approfondie de l'évolution historique des langues d'un groupe linguistique).

Cette connaissance permet de comprendre, de calculer mentalement ou avec une machine, la forme des mots dans une autre langue du groupe. Nous montrerons ci-dessous les équivalences que l'on peut faire entre le tchèque et le slovaque, sur la base de la classification historique, tirée de l'ouvrage de Lamprecht, Šlosar et Bauer :

régularités entre tchèque et slovaque

phénomène phonologique concerné	***slovaque***	***tchèque***	*traduction*
Evolution du « r » mouillé en « ř»	reč	řeč	*« discours »*
Transformations de « a » en « ě » et	ovca	ovce	*« mouton »*
de « u » en « i » derrière une consonne molle	pľúca	plíce	*« poumons »*
Dépalatalisation	hovoriť	hovořit	*« parler »*
Transformation de « ú » en « ou »	súd	soud	*« tribunal »*
Contraction de « ie » en « i »	miera	míra	*« mesure »*
Contraction de « uo » issu de « ó » vers « ů » en tchèque et « ô » en slovaque	pôvod	původ	*« origine »*
Changement de « aj » vers « ej » au sein d'une syllabe	naj-	nej-	*préfixe du superlatif*

Cette connaissance historique fait partie du système linguistique. En termes d'enseignement, c'est la dimension la plus large du système qu'il convient d'enseigner et d'offrir à l'aide de divers outils dont des outils de présentation. Nous montrerons la présentation de la déclinaison substantivale à partir du système qui comprend 7 cas dont le vocatif, un masculin avec 3 sous-genres, un féminin et un neutre et trois nombres : singulier, duel et pluriel. Il est important de faire apparaître le duel qui est constitutif des langues slaves et demeure dans bon nombre d'entre elles, p. ex. haut-sorabe, bas-sorabe, slovène. Dans des langues comme le tchèque, la présence du duel permet d'appréhender globalement les quelques termes qui l'utilisent encore et qui possèdent en plus un pluriel appliqué à d'autres significations que celles afférentes à des organes ou parties du corps.

Nous présentons ci-dessous l'exemple de « ćelo » en haut-sorabe (*veau*) et en tchèque « tele ». Ce sont des mots qui ont gardé leur déclinaison consonantique avec la présence d'un augment. L'application de nos connaissances de grammaire historique permettent de supposer que le « t » de « tele » a été dépalatalisé étant donné que « ć » est en sorabe et en polonais la palatalisation de « t », ce qui correspond au « t' » tchèque. De manière identique, on perçoit que l'augment (dont on apprend par le haut-sorabe qu'il est identique au singulier et au duel) se termine à ces nombres par une consonne palatalisée.

Flexion des substantifs

SUBST — 14

tele	singulier	duel	pluriel
nominatif	tel e		tel at a
génitif	tel et e		tel at
datif	tel et i		tel at ům
accusatif	tel e		tel at a
vocatif	tel e		tel at a
locatif	tel et i		tel at ech
instrumental	tel et em		tel at y

tchèque Patrice POGNAN

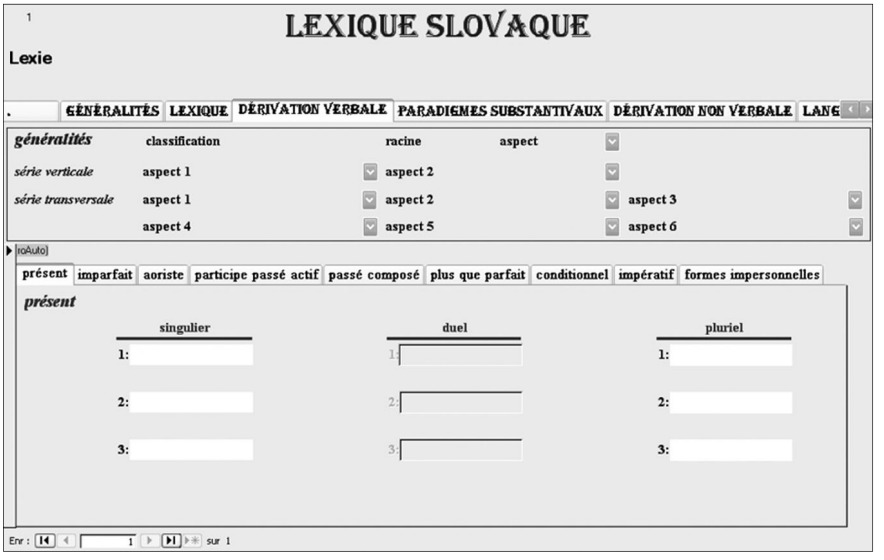
polonais	bs-sorabe	ht-sorabe	tchèque	slovaque	slovène	croate
√ Sm	√ Sm	√ Sm	√ Sm	√ Sm	√ Sm	√ Sm

Dans le même esprit, nous lançons un projet de bases de données de langues slaves de l'Ouest. Ces bases, lorsqu'elles seront consultables publiquement, constitueront également un outil d'apprentissage des langues du groupe. Dans un stade antérieur, c'est un remarquable outil pour faire une analyse détaillée des langues du groupe, tout en ayant la matière pour réaliser des lexiques et dictionnaires de ces langues.

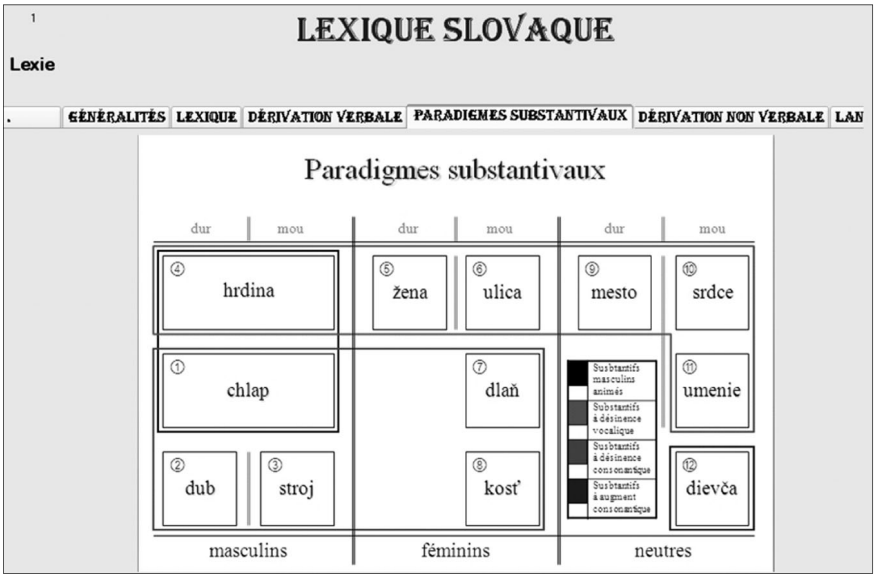
Le projet est parti d'une série de bases de données ayant des fins lexicologiques, en particulier la base de données du slovaque, réalisée pour la constitution des lexiques de la méthode d'apprentissage du slovaque dans le cadre du projet européen ALPCU « découvrir et pratiquer le slovaque », publiée par l'Asiathèque, mais aussi quelques autres : haut-sorabe, berbère, prototype pour le turc. La définition d'une base de données étant chronophage, il a semblé intéressant de définir un prototype général et unique. Cependant, la complexité des systèmes verbaux (par exemple de l'albanais) fait que nous avons généré quelques sous-types découlant du prototype général.

Actuellement les bases de données pour 4 parlers berbères (les 3 marocains et le kabyle) sont prêtes, ainsi que celles destinées au haut-sorabe, au slovaque et au tchèque. Le modèle réalisé pour l'albanais en collaboration avec Klara Lagji est quasiment terminé.

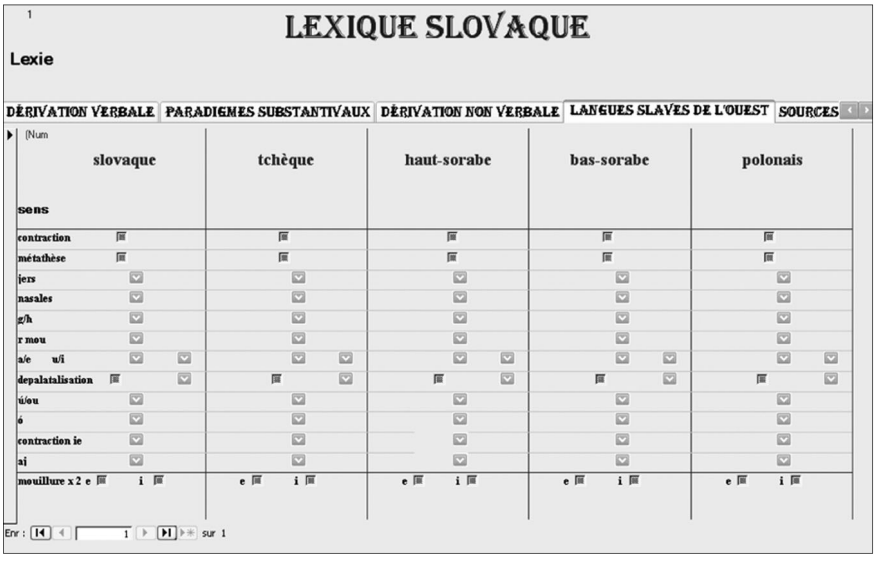
Nous prendrons la base slovaque comme exemple :



Ici aussi, la langue est présentée par rapport au système qui comprend le duel, mais pour le slovaque le duel est montré, mais n'est pas accessible. De même, l'imparfait, l'aoriste et le plus-que-parfait sont présents, mais non activés.



Le fait marquant en ce qui concerne ce dernier onglet, c'est la possibilité d'avoir accès au duel comme nous l'avons dit précédemment.



Mais ce qui est réellement important dans ce projet, c'est la possibilité de comparer les langues du groupe slave de l'Ouest entre elles, deux à deux. Nous pourrons noter tous les changements intervenus entre ces langues par la prise en compte des phénomènes historiques déjà présentés. Cela permettra, en particulier, de juger

de l'importance des phénomènes sur l'ensemble du lexique d'une langue. Typiquement, nous souhaiterions savoir quelle est l'ampleur de la métathèse en haut-sorabe et s'il y a une explication au fait de ne pas produire la métathèse dans certains cas.

En conclusion, nous avons présenté d'une part un système d'analyse automatique du tchèque et d'autre part, quelques outils dont des bases de données qui peuvent répondre tant aux besoins de la pédagogie que profiter à la connaissance scientifique des langues. Dans ces activités qui pourraient sembler éloignées, est toujours présente la composante de la grammaire historique dont l'importance est considérable.

L'ensemble des connaissances réunies sera utile tant à l'apprentissage de langues qu'à la préparation de systèmes d'analyse de celles-ci.

Bibliographie :

Bielec Dana (1998), *Polish. An Essential Grammar Routledge*, Londres, New York.
Blanche-Benvéniste Claire (1997), *Eurom4 : méthode d'enseignement de quatre langues romanes.* Firenze : A. Valii et alii, Nuova Italia Editrice.
Caduc Eveline & Castagne Eric (2001), *Pour une modélisation de l'apprentissage simultané de plusieurs langues apparentées ou voisines.* Nice : Université de Nice-Sophia Antipolis.
Castagne Eric (Ed.) (2004), *Intercompréhension et inférences.* Reims : Presses universitaires de Reims.
Castagne Eric, *Programme InterCompréhension Européenne (ICE).* Site internet : http//:logatome.eu.
Dobrovský Josef (1951), *Dějiny české řeči a literatury*, Československý spisovatel, Praha.
Havránek Bohuslav & Jedlička Alois (1970), *Česká mluvnice*, Státní pedagogické nakladatelství, Praha.
Král' Ábel (1996), *Pravidlá slovenskej výslovnosti*, Slovenské Pedagogické Nakladatel'stvo, Bratislava.
Kurz Josef (1967), *Jazyk staroslověnský*, Filosofická Fakulta University Karlovy, Státní pedagogické nakladatelství, Praha.
Lamprecht Arnošt, Šlosar Dušan & Bauer Jaroslav (1986), *Historická mluvnice češtiny*, SPN, Praha.
Mareš František Václav (2000), *Cyrilometodějská tradice a slavistika*, Torst, Praha.
Marvan Jiří (2000), *Jazykové milénium. Slovanská kontrakce a její český zdroj*, Academia, Praha.
Mazon André (1952), *Grammaire de la langue tchèque*, Institut d'Etudes Slaves, Paris.
Mistrík Jozef (1983), *Moderná slovenčina*, Slovenské Pedagogické Nakladatel'stvo, Bratislava.
Slodzian Monique & Souillot Jacques (Eds.) (1997), *Compréhension multilingue en Europe.* Paris : sous les auspices de la Communauté européenne, CNRS & INALCO.
Starosta Manfred (1999), *Dolnoserbsko-nimski słownik*, Nakładnistwo Domowina, Bydyšyn.
Šewc-Schuster Hinc (1984), *Gramatika hornjoserbskeje rěče*, Nakładnistwo Domowina, Budyšin.