Description formelle et diagnostic automatique des erreurs en langue étrangère : quelques perspectives pour les outils d'ELAO

Ivan Šmilauer

Ivan Šmilauer
Université Paris IV, INALCO – LaLIC-CERTAL

Description formelle et diagnostic automatique des erreurs en langue étrangère : quelques perspectives pour les outils ELAO

Résumé
Nous présentons les possibilités des nouvelles technologies et du traitement automatique des langues (TAL) appliquées aux outils d'enseignement des langues assisté par ordinateur (ELAO). Sur l'exemple de la plateforme www.cetlef.fr, *nous montrons que le développement d'un diagnostic automatique des erreurs met en jeu les deux aspects majeurs d'une normalisation des données linguistiques : d'une part, la normalisation du métalangage descriptif et classificatoire des données exigé par le traitement informatique ; d'autre part, la normalisation d'un texte mal formé par rapport à une norme établie. Le diagnostic présenté permet de générer une réponse interactive aux activités langagières des apprenants. Les résultats obtenus dans un cadre expérimental ciblé sur des exercices de flexion nominale tchèque offrent un bon point de départ pour la prise en compte des niveaux linguistiques supérieurs.*

---

1. ELAO et normalisation

L'enseignement des langues assisté par ordinateur (ELAO) est un domaine pluridisciplinaire nécessitant un mariage de compétences des linguistes, des pédagogues et des informaticiens[1]. Les outils ELAO doivent faire leurs preuves en confrontation avec les besoins réels d'un apprenant de langue – leur « raison d'être » ainsi que leur meilleur juge. Dans la perspective du croisement des nouvelles technologies et des sciences du langage, ce domaine est un excellent exemple de l'interaction entre les deux mondes.

Dans notre communication, nous abordons cette question par le biais de la normalisation linguistique qui fut choisie comme un des points d'articulation du colloque ASL'09. Après une introduction sur les questions de normalisation, nous allons présenter les possibilités de l'intégration des techniques du traitement automatique des langues (TAL) dans les outils ELAO et nous l'illustrerons par une réalisation concrète qui est un dispositif destiné aux apprenants de la langue tchèque avec un diagnostic automatique des erreurs de déclinaison.

1.1 Les aspects de la normalisation

Dans sa définition classique[2], la normalisation signifie la création, la fixation et l'application d'une norme ; la normalisation d'un objet consiste à le rendre conforme à une norme donnée. Considérée sous un angle technique, une norme est un ensemble explicite de propriétés et de règles concernant un objet qui, grâce à la norme imposée, peut être manipulé et reproduit par des sujets indépendants. L'exemple typique sont les normes technologiques industrielles, informatiques, éditoriales, etc. Ces normes sont la condition nécessaire d'une collaboration entre des sujets différents et le gage d'un développement continu qui puisse bénéficier des acquis existants. Dans la perspective linguistique, la norme, en tant que terme issu de la dichotomie *usage – norme*, devrait

[1] Pour une vue succincte du domaine, voir par ex. Levy (1997), pour ses différents aspects didactiques, voir Demaizière (2007). Dans le milieu anglophone, voir notamment les revues *CALICO* (https://calico.org/), *ReCALL* (http://www.eurocall-languages.org/recall/). Pour la production francophone, voir la revue *ALSIC* (http://www.alsic.org).
[2] Voir *Lexis*. Larousse (2002).

être comprise plutôt comme un consensus collectif sur ce qui est, dans l'usage d'une langue, considéré par ses locuteurs comme « normal »[3]. A priori, une telle norme n'est pas imposée de l'extérieur, bien que les initiatives de normalisation linguistique existent dans la plupart des sociétés modernes, en se manifestant notamment par l'élaboration des règles d'orthographe et d'orthoépie, des grammaires normatives et des recommandations stylistiques.

Dans le contexte du questionnement sur la place des nouvelles technologies informatiques dans les sciences du langage et plus spécifiquement, sur leur intégration dans les outils ELAO, la question de normalisation peut être donc appréhendée de deux positions différentes. Premièrement, il est possible de s'interroger sur les normes imposées de l'extérieur de la langue qui déterminent la matérialité des documents textuels (comment les textes sont-ils inscrits sur un support numérique) et le métalangage descriptif (comment les textes sont-ils décrits). En second lieu, nous pouvons nous questionner sur les caractéristiques inhérentes des textes et leur relation par rapport à la norme comprise au sens linguistique (comment les textes sont-ils écrits).

1.2 Matérialité des documents et métalangage descriptif

Bien qu'il s'agisse d'un problème principalement technique, la définition des normes pour le codage des documents textuels ne se passe pas sans l'intervention des linguistes, notamment pour l'élaboration de la norme Unicode[4]. C'est à ce niveau, nécessaire pour tout traitement numérique des données linguistiques, que les normes sont les plus restrictives et les mieux observées, malgré quelque incompatibilités subsistantes dues à la concurrence industrielle (existence des normes parallèles ISO, MS Windows et Apple Mac). La norme Unicode veut remédier à cette situation et elle prétend imposer un seul format de codage de caractères à toutes les langues du monde. A part quelques problèmes spécifiques liés au traitement des documents multilingues, qui peuvent être résolus grâce à la norme Unicode, il n'y a pas de point particulier à ce niveau qui poserait de problème dans le développement des outils ELAO.

Concernant le métalangage, les possibilités d'enrichissement des documents textuels par des annotations variées sont très larges et vont de la description des facteurs situationnels déterminant la genèse du texte (auteur, médium, lieu et moment de l'énonciation, etc.) jusqu'à l'annotation de ses différentes unités et structures linguistiques (annotation morphologique, syntaxique, sémantique, etc.). Ces informations peuvent être exploitées par un humain (par exemple pour des recherches de corpus paramétrées par une information grammaticale) et par une machine lors du traitement automatique.

La volonté de créer un métalangage descriptif universel rejoint l'éternelle ambition des linguistes de décrire une langue à l'aide d'étiquettes simples, non ambiguës et qui seraient attribuées à l'aide de règles, si possible, sans exceptions. La complexité des données linguistiques et la variété des approches théoriques rendent la définition et l'acceptation d'une norme unique de description difficile, voir impossible. Il est donc habituel que les normes d'annotation s'établissent individuellement au sein des différents projets scientifiques suivant les divers fondements théoriques et elles doivent se plier souvent aux exigences et limitations imposées par un outil

[3] Voir *Dictionnaire de linguistique.* Larousse 2002, p. 330.
[4] Voir http://unicode.org/.

concret[5]. A condition que les règles pour l'attribution des étiquettes soient explicitement définies, il ne devrait pas, a priori, être difficile de « traduire » un jeu d'étiquettes dans un autre, basé sur une approche différente, mais établi avec autant de soin que la première.

Comme nous allons le montrer, dans le cadre d'une application ELAO intégrant le TAL pour le diagnostic des erreurs, la question de normalisation doit être abordée à deux reprises. Premièrement, en relation avec l'annotation des textes de la langue cible ; deuxièmement, au cours de la définition d'une description des erreurs dans les productions des apprenants.

2.3 Norme linguistique

Dans le cas où nous comprenons le terme « norme » dans son acceptation linguistique et nous considérons la normalisation d'un texte comme son remaniement fidèle à un consensus collectif des locuteurs d'une langue donnée, nous avons le droit de considérer la correction des textes comportant des déviations par rapport à cette norme également comme une sorte de normalisation. Le degré de l'intervention peut être variable, allant des réajustements stylistiques minimes dans un énoncé d'un locuteur natif jusqu'à l'élimination des structures agrammaticales dans des productions des apprenants d'une langue étrangère.

Avant d'être un sujet important pour les recherches en TAL, la correction des énoncés est une question complexe du point de vue linguistique : quoi et à quel niveau peut être considéré comme correct ou incorrect, grammatical ou agrammatical, acceptable ou inacceptable, approprié ou inapproprié, etc. A côté des correcteurs d'orthographe et les vérificateurs de grammaires, implémentés actuellement dans des logiciels destinés au grand public, les problèmes liés à la correction sont sensiblement plus difficiles à résoudre lors du développement des outils ELAO qui ont l'ambition d'analyser les productions en langue étrangère, nécessairement imparfaites et contenant des erreurs spécifiques.

2. Intégration des outils de TAL dans ELAO

Les premiers essais pour améliorer les outils ELAO par l'emploi du TAL datent du début des années 80 et ils se sont produits parallèlement à la recherche sur les tuteurs intelligents en intelligence artificielle, impliquant la modélisation de l'apprenant et du processus de l'apprentissage, voir par ex. Bruillard (1997), Dodigovic (2005). Ce sont les travaux dans ce domaine qui ont emmené un qualificatif quelque peu prétentieux : les outils d'enseignement / d'apprentissage des langues « intelligemment » assisté par ordinateur[6]. Sans qu'il existe une frontière nette entre ceux qui peuvent être considérés comme « intelligents » et ceux qui ne le sont pas, il est possible de dire que les premiers ne se limitent pas seulement à la reproduction des contenus et à des activités

[5] Voir par ex. les définitions de l'annotation morphologique et syntaxique dans Prague Dependency Treebank, fondée sur les principes de la syntaxe de dépendance (http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch02.html) ou pour Penn Treebank, présenté par ex. chez Jurafsky (2009), p. 165, fondé sur la syntaxe syntagmatique.

[6] Quant à l'appellation du domaine, les outils et la recherche adjacente sont désignés dans le milieu anglophone par le sigle ICALL (Intelligent Computer Assisted Language Learning), voir par ex. Heift & Schulze (2007), Meurers (2009). Dans le milieu francophone, on utilise le plus souvent les termes ELAO / ALAO (Enseignement / Apprentissage des Langues Assistés par Ordinateur) intégrant le TAL, voir par ex. Hamel & Vandeventer (1998), Antoniadis et al. (2005). La discipline couvrant ce domaine est appelée traditionnellement EIAO (Enseignement Intelligemment Assisté par Ordinateur).

traditionnelles sur le support électronique des nouveaux médias (hypertexte, audio, vidéo, communication en ligne, réalité virtuelle, etc.) mais qu'ils essayent également de faire profiter les apprenants des derniers acquis en TAL pour rendre les outils ELAO plus riche. En principe, pour chacune des techniques et des applications de TAL, il est possible de trouver une façon comment l'intégrer dans un outil ELAO[7].

Pour faciliter l'orientation du lecteur, nous présentons d'abord une revue très simplifiée de ces techniques et outils[8]. Leur complexité monte avec la complexité linguistique des tâches qui leur sont imposées – les instruments de bas niveau sont employés dans les applications plus sophistiquées, souvent de caractère modulaire. En allant de l'analyse des chaînes de caractères dont les textes sont composés vers l'analyse de leur structure, les outils utilisés sont les tokenizers (segmentation de la chaîne en unités), les lemmatiseurs et les taggers (analyse morphologique) et les parseurs (analyse syntaxique). Des techniques correspondantes sont utilisées également dans le sens de la génération automatiques des textes. Ces outils peuvent être employés dans des applications du niveau supérieur comme des correcteurs d'orthographe, de grammaire et de style, la traduction automatique, l'analyse sémantique et pragmatique des textes, la recherche d'information, le résumé automatique, etc. Un produit spécifique lié au TAL sont les corpus électroniques, monolingues ou multilingues, et les instruments destinés à leur exploitation. En fonction des techniques choisies, de nombreux outils cités impliquent la compilation des grammaires et des dictionnaires électroniques exploitables d'une part par une autre application, d'autre part, dotées d'une interface appropriée, par un humain. Les difficultés liées au traitement du signal sonore s'ajoutent au domaine spécifique de la reconnaissance et synthèse vocale.

Globalement, il est possible de distinguer deux façons fondamentales d'intégrer ces outils dans ELAO : (1) enrichissement des ressources pédagogiques en langue étrangère ; (2) le traitement des productions langagières des apprenants.

2.1. Enrichissement des ressources

L'enrichissement des ressources mises à la disposition de l'apprenant peut être à son tour effectué de deux manières : (1) emploi des outils de bas niveau (taggers, parseurs, etc.) pour une annotation linguistique des textes pédagogiques (lemmatisation et ajout des étiquettes morphologiques, lexicales, syntaxiques ou sémantiques) ; (2) intégration des outils de haut niveau (traduction automatique, dictionnaires, corpus, traitement du signal sonore) pour diversifier l'exposition de l'apprenant à la langue cible.

Grâce à l'enrichissement du premier type, un apprenant peut consulter les propriétés linguistiques des différentes unités linguistiques faisant partie des textes en langue étrangère : le lemme d'un certain lexème, ses catégories grammaticales et lexicales, son type de flexion (le cas échéant), sa fonction syntaxique, etc. La condition nécessaire de ce type d'enrichissement est la fiabilité de l'annotation. Or, un taux d'erreurs constant est produit par les analyseurs actuels[9]. Une correction manuelle est donc nécessaire afin de ne pas induire l'apprenant en erreur. Au sein d'un

[7] Pour les différentes réalisations concrètes, voir notamment Nerbonne (2003) ou Heift & Schulze (2007).
[8] Une revue quasi-exhaustive du domaine est présentée par ex. dans Jurafsky & Martin (2009).
[9] Pour le tchèque, le taux d'erreurs des taggers existants est rarement inférieur à 5 %, (en moyenne, un mot sur vingt dans un texte est mal analysé), voir Spoustová (2008). Pour l'anglais, ce plafond est établi autour de 3 %, voir Jurafsky (2009), p. 195.

outil ELAO, la consultation de ces informations peut être effectuée d'une manière classique, semblable à la consultation d'un dictionnaire (consulter un élément et ses propriétés), mais il est également possible de profiter du médium électronique pour une visualisation plus attractive de ces données (par exemple par la coloration ou le changement de taille de police des unités en fonctions de leurs attributs ou la mise en relation graphique des unités liées). Du point de vue de l'architecture du dispositif, l'annotation des textes permet de procéder à la génération automatique des exercices ciblés sur un certain phénomène linguistique (par ex. sur l'emploi des prépositions, sur l'accord dans le groupe nominal, etc.) ou à l'évaluation automatique de la difficulté des textes (lexicale et grammaticale) qui permettrait d'accommoder leur choix aux compétences de l'apprenant ou à la structuration didactique de la méthode.

Les modalités de l'emploi des outils plus sophistiqués reflète leur diversité. Avec l'état de l'art actuel, intégrer une grammaire ou un dictionnaire électronique et des ressources basés sur corpus est mieux envisageable que de se servir des outils existants pour la traduction, le résumé automatique ou des modules de dialogue homme – machine qui n'atteignent pas une fiabilité satisfaisante. Le premier type d'outil peut être utilisé pour la présentation des équivalents en langue maternelle de l'apprenant, pour la génération des formes fléchies d'un lexème, pour la consultation de ses occurrences dans des contextes authentiques, pour la consultation des listes de collocations les plus fréquentes, pour la consultation des lexiques structurés en réseau du type WordNet, etc.

De nombreux outil ELAO d'aujourd'hui contiennent des modules de reconnaissance vocale, cependant, ils ne sont pas, dans la plupart des cas, utilisés pour une analyse phonologique qui permettrait par la suite de traiter la production orale de l'apprenant comme un énoncé écrit, mais ils se limitent uniquement à l'analyse phonétique, servant à la vérification de la prononciation par la comparaison avec des modèles préenregistrés.

2.1. Traitement des productions de l'apprenant

La seconde possibilité comment intégrer des techniques du TAL dans ELAO est de s'en servir pour l'analyse des énoncés produits en langue étrangère par les apprenants au sein des différentes activités pédagogiques, notamment des exercices grammaticaux, de traduction et de rédaction. Les résultats de cette analyse peuvent être profitables aux apprenants par le biais d'une rétroaction appropriée à leurs productions (la correction et le diagnostic de leurs erreurs), mais également aux chercheurs en acquisition d'une langue étrangère grâce à l'ajout des informations métalinguistiques, utiles pour l'exploitation automatisée des corpus de productions.

L'emploi des outils génériques du TAL pour l'enrichissement des ressources pédagogiques ne met pas en cause leur fondement théorique et fonctionnel – les textes destinés à l'apprenant ne sont qu'un type spécifique d'objet pour lequel ils ont été conçus et les problèmes qui se posent pour ce type d'intégration ne sont pas tant d'ordre technique que d'ordre pédagogique. La situation est différente si nous considérons l'emploi du TAL pour le traitement des productions des apprenants, reflétant leur interlangue et présentant de nombreuses spécificités : elles peuvent contenir des erreurs à tous les niveaux (orthographe, morphologie, syntaxe, sémantique, pragmatique), leur contenu est souvent construit à partir d'un ensemble d'unités limité et de structures linguistiques dont le choix et l'usage est influencé par le niveau de l'apprenant et par des stratégies d'évitement, voir par ex. Gaonac'h (1991). Bien que le caractère restreint de ces textes puisse être un facteur

favorable pour un traitement automatique, la présence des erreurs le rend plus complexe par rapport au traitement des textes en langue standard.

Comme le montrent, sans surprise, les expériences avec le traitement automatique des corpus d'apprenants, les outils de TAL génériques atteignent des pourcentages de réussite encore plus bas que pour les textes en langue standard, voir par ex. Granger et al. (2001), Johannessen et al. (2002). Si un certain nombre d'erreurs produites par l'analyse automatique peut être toléré dans les applications destinées à un locuteur natif, elles sont littéralement néfastes pour un apprenant de langue quant à la correction de ses propres erreurs par rapport à la norme de la langue cible. Ce constat a donné naissance à des recherches en TAL dont l'objectif est de pallier cette imperfection et de procéder à des traitements qui permettraient non seulement de corriger les erreurs mais également de déterminer leurs causes (diagnostic) afin de produire un message de rétroaction destiné à l'apprenant. Une collaboration avec des experts en acquisition de langue étrangère, et plus spécifiquement, en analyse des erreurs est nécessaire, afin de modéliser correctement l'activité linguistique et cognitive de l'apprenant résultant en productions erronées, voir notamment Schulze (2008).

Dans les quinze dernières années, les travaux sur le diagnostic des erreurs et une rétroaction qui ne serait pas limitée à la simple constatation « *correct / incorrect* », sont le fil principal des recherche en ELAO intégrant le TAL, voir notamment Heift & Schulze (2003, 2007) ou Meurers (2009). En considérant les outils ELAO comme un complément d'enseignement en présentiel, l'accent mis sur cette problématique est justifié par une durée relativement courte qui peut être consacrée dans les cours de langues à la correction individuelle. Dans les outils destinés aux autodidactes, la rétroaction automatique est le seul retour qu'un apprenant puisse avoir sur ses productions.

Du point de vue de la complexité linguistique des données traitées, la correction et le diagnostic des erreurs peuvent être effectués sur les productions issues de deux types de tâches : (1) tâches fermées et (2) tâches ouvertes.

Les tâches fermées sont celles où le nombre de réponses correctes est limité. Il s'agît le plus souvent des exercices grammaticaux ou lexicaux (phrase à trous, exercices de transformation), des réponses attendues à des questions précises ou de la traduction des énoncés simples. Le traitement des productions issues de ces tâches est facilité car il est possible de mettre toutes les réponses possibles à la disposition de l'analyseur et de se concentrer uniquement sur les différences qui les distinguent de la production donnée. En fonction des phénomènes linguistiques, on peut utiliser des techniques simples, utilisant uniquement le calcul des différences entre les deux chaînes de caractères, voir par ex. Desmets (2006), ou des techniques plus sophistiquées qui sont employées pour l'analyse des structures morphologiques et syntaxiques des productions issues des tâches ouvertes, voir nottament Heift & Schulze (2007).

Les taches ouvertes (ou libres) sont des tâches rédactionnelles dans lesquelles l'apprenant produit des énoncés contenant des structures complexes qui peuvent aller d'une phrase simple à plusieurs paragraphes. Leur contenu peut être éventuellement réduit par une restriction sur le sujet ou le genre de la rédaction mais il est en général imprévisible du point de vue des unités et structures linguistiques qui y seront employées. Du point de vue technique, il existe deux méthodes

pour le diagnostic et la correction de ce type spécifique de textes. La première consiste en modification des algorithmes de d'analyse syntaxique dans les parseurs existant afin qu'ils puissent analyser des structures agrammaticales. Celles-ci peuvent être interprétées en repérant les contraintes syntaxiques qui ont été enfreintes dans les règles de la grammaire du parseur, afin que l'analyse d'une phrase erronée puisse aboutir à une structure syntaxique acceptable. Cette technique, appelée le « relâchement de contraintes », a été utilisée par exemple pour le diagnostic des erreurs en français au sein du projet FreeText, voir L'Haire & Faltin (2003). La deuxième méthode consiste en extension de la grammaire du parseur par des règles spécifiques – les « mal rules » – inspirés par les erreurs les plus courantes repérées dans des corpus de productions libres et qui acceptent les structures erronées, voir par ex. l'article de Schneider & McCoy (1998) traitant de l'anglais comme la langue cible au sein du projet ICICLE. Le diagnostic des erreurs est effectué par chacune de ses méthode soit sur la base de la contrainte relâchée, soit en fonction de la règle illicite appliquée pour l'analyse de la structure erronée.

Bien que l'objectif ultime du diagnostic automatique soit le traitement des productions libres, il est encore loin d'être atteint à l'heure actuelle. Les analyseurs adaptés affrontent non seulement les problèmes classiques liés à l'ambiguïté des unités et des structures linguistiques mais de plus, l'introduction du relâchement ou des « mal-rules » conduit à des résultats assez bruités, avec des erreurs signalées dans des constructions tout à fait correctes. La stratégie la plus convenable pour le moment semble être de se limiter à des techniques bien maîtrisées (correction orthographique et analyse morphologique) et de les employer d'une manière ciblée pour un diagnostic fiable, traitant uniquement des phénomènes linguistiques simples et bien définis, voir par ex. Kraif et al. (2004). Il apparaît également qu'il est plus approprié de se limiter d'abord à la résolution des problèmes liés au diagnostic des productions issues des tâches fermées, car cette restriction du cadre expérimental permet de mieux identifier et contrôler les différents points critiques et préparer le terrain pour des analyses plus complexes.

3. Diagnostic des erreurs sur CETLEF.fr

Afin d'illustrer cette approche « pragmatique », nous présentons ici le diagnostic des erreurs de déclinaison tchèque au sein de l'outil CETLEF.fr[10]. Il s'agit d'une application web dynamique disponible publiquement sur http://www.cetlef.fr à partir de juin 2008. Toute personne intéressée peut utiliser le dispositif, à condition qu'elle remplisse un formulaire d'entrée pour donner quelques informations (âge, durée de l'apprentissage du tchèque, autres langues maîtrisées, etc.) qui peuvent être utilisées lors de l'analyse des productions, stockées intégralement dans la base de données.

Le tchèque, une langue slave occidentale parlée par environ 11 millions de locuteurs, possède une flexion nominale très riche : sept cas (nominatif, génitif, datif, accusatif, vocatif, locatif, instrumental), deux nombres (singulier, pluriel), quatre genres (masculin animé et inanimé, neutre, féminin), 14 paradigmes de déclinaison nominale avec de nombreux sous-types, des exceptions et des alternances morphématiques vocaliques ou consonantiques effectuées sur le radical en fonction de la désinence casuelle. Cette complexité représente non seulement une difficulté pour l'apprentissage du point de vue d'un apprenant, mais également un défi pour le

[10] CETLEF – *Connaître, Comprendre et Corriger les Erreurs en Tchèque Langue Étrangère pour les Francophone*s, voir Šmilauer (2008).

traitement automatique, il est donc évident que l'aspect morphologique des productions s'impose comme le premier problème à résoudre.

3.1 Description du dispositif

CETLEF.fr propose des exercices de type phrase à trous où la tâche de l'apprenant est de décliner un substantif, un adjectif ou un pronom en fonction de son contexte syntaxique. L'exemple d'une tâche avec la production de l'apprenant est représenté sur la figure 1.

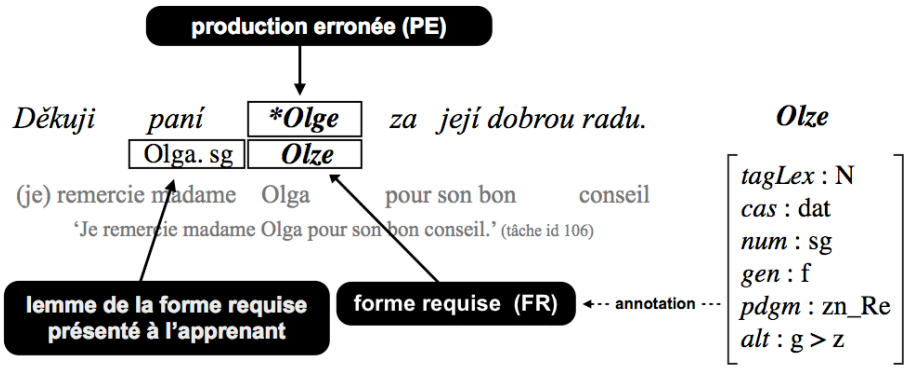

Figure 1. – Description d'une tâche de déclinaison et annotation de la forme requise.

Il peut y avoir au maximum 3 formes requises (FR) dans une tâche et elles sont saisies manuellement par l'auteur sur une plateforme dédiée à la conception des exercices. Si la production de l'apprenant n'est égale à aucune des FR, nous l'appelons production erronée (PE). Du point de vue computationnel, le cadre restreint d'une tâche fermée permet de nous concentrer d'abord sur les problèmes liés au traitement de l'orthographe (aspect graphique des mots) et de la morphologie (structure des mots) avant d'accéder, dans les travaux futurs, au niveau de la syntaxe (structure des phrases). Du point de vue de l'apprenant, une telle tâche peut être considérée comme artificielle et la question de savoir à quel point cette activité témoigne de sa maîtrise effective de la langue cible, est tout à fait légitime. Cependant, la logique de résolution des problèmes simples avant d'aborder des traitements des plus complexes justifie cette restriction. De plus, les exercices grammaticaux de ce type sont très courants dans l'enseignement des langues.

Chaque FR est annotée à l'aide d'un analyseur morphologique rudimentaire qui spécifie la catégorie lexicale, le cas, le nombre, le genre, le type de déclinaison et une éventuelle alternance morphématique (voir figure 1). Afin de ne garder qu'une seule étiquette pour chaque FR, le choix parmi plusieurs annotations possibles, dues à l'homonymie des désinences casuelles, est effectué manuellement par l'auteur. L'annotation des FR a un triple objectif : (1) elle est nécessaire pour le diagnostic des erreurs, (2) elle peut être utilisée pour l'exploitation des productions dans la base de données et (3) elle est employée pour la visualisation de l'information grammaticale destinée à l'apprenant.

A l'heure actuelle, le dispositif contient une batterie d'exercices qui couvrent toute la déclinaison du tchèque (avec la difficulté montante qui suit la présentation des différents cas dans les méthodes d'enseignement traditionnelles) et une autre, destinée aux débutants qui sert comme complément aux cours de la langue tchèque à l'INALCO pendant l'année 2009–2010. Le parcours

de l'apprenant à travers les exercices est donné d'une manière fixe, ce qui assure une consistance de son avancement en accord avec les programmes pédagogiques. Après avoir terminé un exercice, l'apprenant peut consulter sa correction avec le message de diagnostic pour les PE et les propriétés morphologiques des FE (voir figure 2).

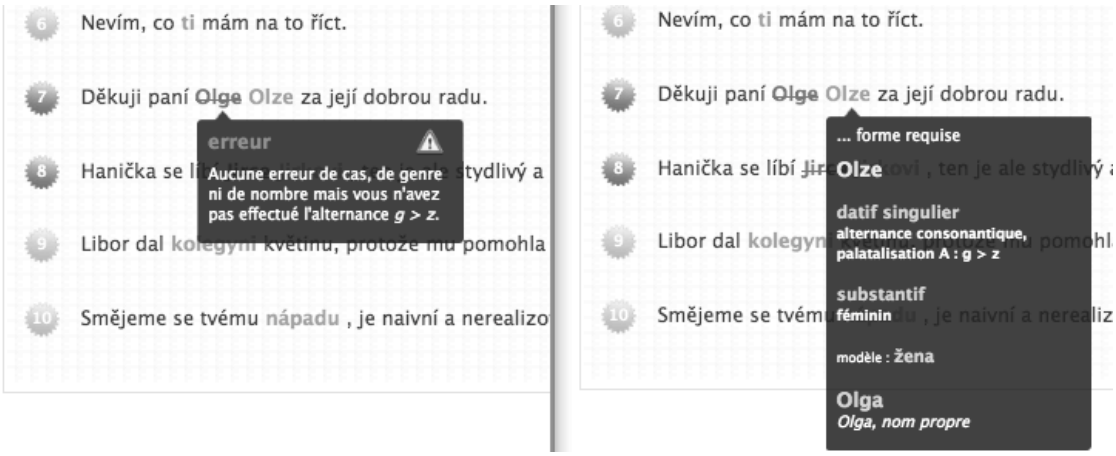

Figure 2. – Le diagnostic de la PE et la visualisation de l'annotation de la FR.

Au fur et à mesure que l'apprenant avance dans les exercices, une base de données personnelle, propre à chaque apprenant, se remplit par les éléments lexicaux et morphologiques rencontrés dans les exercices (différents lexèmes, formes casuelles, types de déclinaison et alternances) qui peuvent être consultés comme une sorte de dictionnaire électronique.

3.2 Fonctionnement du diagnostic

La conception du diagnostic a été inspirée par des travaux en analyse des erreurs – méthode utilisée dans les recherches sur l'acquisition d'une langue étrangère depuis les années 70. Ces analyses ont rendu compte des différents types d'erreurs, allant des interférences avec la langue maternelle aux erreurs dues à la surgénéralisation et le non-respect des idiosyncrasies, voir par ex. Porquier (1977). Cette question est d'actualité également dans les travaux menés dans la perspective psycho-computationnelle, voir par ex. Pirelli (2007), qui tentent de modéliser l'acquisition et le traitement cognitif de la morphologie à l'aide des formalismes explicites.

En nous fondant sur les acquis intérieurs, nous partons de l'hypothèse qu'une PE dans une tâche de déclinaison n'est pas une forme aléatoire mais calculable – elle est le résultat d'une activité succombant à des règles d'ordre cognitif et elle peut être reproduite artificiellement d'une manière contrôlée. Simplement dit, afin de comprendre l'erreur de l'apprenant, nous essayons de la produire automatiquement. Ainsi nous supposons qu'une PE peut être générée à partir du lemme de la FR à l'aide de (1) l'ajout des désinences casuelles inappropriées, (2) par le non-respect des alternances morphématiques obligatoires et (3) par la violation des règles de graphie (diacritique, casse). L'avantage de cette technique réside dans le fait qu'elle permet d'interpréter également les formes

inexistantes dans la langue cible mais qui sont tout à fait prévisibles et « logiques » (par ex. *chevals*, *prendons* en français)[11].

Du point de vue technique, le diagnostic d'une PE est effectué par sa comparaison avec des formes hypothétiques (FH), générées à partir du radical de la FR (voir figure 3).

FR = '**olze**' PE = '**olge**'

FH n°	radical	désinence	cas	nombre	alternance	FH = PE	interprétation
1	**olg**	**a**	nom	sg	∅	0	∅
2	**olg**	**y**	gén	sg	∅	0	∅
3	**olz**	**e**	dat	sg	g > z	0	∅
4	**olg**	**e**	**dat**	**sg**	**∅**	**1**	**erreur d'alternace**
…							

Figure 3. – La génération des FH et recherche d'une correspondance entre les FH et PE.

Cette génération est possible grâce à un modèle formel de la déclinaison tchèque contenant les paradigmes de déclinaison et les règles d'alternance (voir plus bas). Si une correspondance est trouvée, la différence dans les structures de traits morphologiques de la FH et de la FR est retenue comme une possible interprétation de la PE et une telle PE est appelée erreur morphologique (voir figure 4).

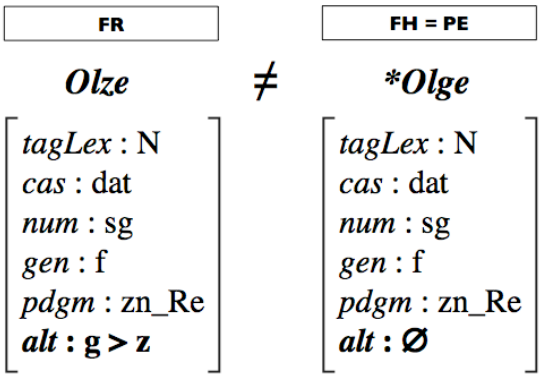

Figure 4. – La comparaison de la FR avec une des FH avec une interprétation 'erreur d'alternance'.

Pour chaque PE, le diagnostic génère un code d'erreur qui est inscrit dans la base de données. Ce code peut être exploité lors des évaluations statistiques des différents types d'erreurs mais il sert principalement pour la rédaction automatique du message de rétroaction destiné à l'apprenant (voir figure 2).

[11] Une approche similaire, basée sur la comparaison entre réponse attendue et réponse donnée, a été proposée dans Kraif & Ponton (2007) dans le cadre du système ExoGen. Un prototype de correcteur fondé sur la prévisibilité des erreurs dans la flexion verbale de l'allemand a été proposé dans Rimrott (2003).

Avant la fin de février 2010, 37 exercices contenant au total 375 tâches ont été publiés sur CETLEF.fr. Depuis l'ouverture du site en juin 2008, 178 apprenants inscrits ont envoyé à la correction 8180 productions dont 2937 (35,9 %) sont des PE. Le nombre de PE diagnostiquées en tant qu'erreur morphologique est 2330 (79,3 % de toutes les PE) ce qui est un résultat encourageant pour l'algorithme du diagnostic ainsi que pour l'hypothèse sur la calculabilité des erreurs. Parmi les PE restantes, 4 % environ ont pu être interprétés par des méthodes non-morphologiques[12] et 16 % des PE restent sans interprétation, c'est-à-dire qu'aucune des formes hypothétiques générées par le diagnostic n'est égale à la PE fournie par l'apprenant.

4. Questions de normalisation

Les problèmes liés à la normalisation des données ont dû être résolus dans deux domaines distincts : (1) développement du modèle de la déclinaison tchèque nécessaire pour l'annotation des FR et pour le diagnostic des PE ; (2) description formelle des PE.

4.1 Modèle de la déclinaison tchèque

Le tchèque est une langue dotée de ressources linguistiques électroniques très riches[13]. Il existe deux modèles formels de la morphologie tchèque avec un format normalisé qui permet son implémentation dans des différentes applications de TAL : le premier est celui utilisé pour le tagger de Hajič (2004), le second est utilisé dans l'application *ajka,* présenté dans Sedláček & Smrž (2001). Cependant, aucun de ces modèles n'a pu être utilisé au sein de CETLEF.fr. En effet, il s'agit d'analyseurs morphologiques qui ont été conçus pour le traitement efficace et rapide de grandes quantités de textes[14]. Avec cet objectif en vue, le traitement des alternances morphématiques qui nécessitent un traitement sensible au contexte, a été résolu d'une manière fonctionnelle au niveau informatique mais peu adéquat pour la génération des formes où les règles d'alternances doivent être enfreintes.

En prenant la PE dans les figures 1, 2 et 3 comme exemple (*olge* au lieu de *olze* avec la segmentation en radical *olg / olz* et *la* désinence *-e*), l'erreur de l'apprenant a été causée par le non-respect de la palatalisation *g* > *z* qui doit se produire régulièrement sur un radical terminé par la consonne *g* si la désinence est *-e*. À cause de leur complexité computationnelle, le modèle de Hajič n'introduit pas les règles pour les alternances et il résout ce problème par l'introduction des paradigmes spécifiques où, dans la représentation formelle, la consonne alternée fait partie de la désinence – la segmentation est donc *ol-ze,* ce qui n'a pas de fondement dans le système de la flexion nominale du tchèque. Pour des alternances vocaliques qui se produisent à l'intérieur du radical (par exemple *stůl* > *stol-u*), le dictionnaire morphologique de ce modèle intègre les

[12] Avant le lancement du diagnostic morphologique, relativement coûteux en ressources informatiques, une série de contrôles simples sur les caractères et les chaînes, sans fondement linguistique, est lancée afin d'identifier des erreurs sans motivation morphologique (par ex. la distance de Levenshtein trop importante, la présence de caractères qui n'appartiennent pas à l'alphabet tchèque, etc.).

[13] Mentionnons notamment le *Corpus National Tchèque* (ČNK), voir Čermák (1997), disponible sur http://ucnk.ff.cuni.cz/, contenant plusieurs centaines de millions de mots, dont une partie annotée morphologiquement. Une autre réalisation importante est le *Prague Dependency Treebank* (PDT), voir Hajič (2005), disponible sur http://ufal.mff.cuni.cz/pdt, qui contient quelque 2 millions de mots avec une annotation morphologique, syntaxique et sémantique et qui est suivi d'une large gamme d'outils pour l'exploitation des données formatées d'après la norme PDT.

[14] Le tagger de Hajič a été utilisé pour l'annotation morphologique de ČNK et de PDT.

paradigmes entiers de chaque lexème en question. Dans *ajka,* le problème des alternances est contourné par l'ajout des segments supplémentaires entre la désinence casuelle et le reste du radical sans les consonnes ou voyelles alternées. La segmentation de *olze* est donc *ol-z-e* et un nouveau paradigme est introduit pour rendre compte de la flexion de tous les substantifs avec cette alternance.

La solution de *ajka* pourrait être mieux adaptée pour nos besoins, nous avons cependant opté pour la conception d'un modèle spécifique où les paradigmes seraient définis uniquement par les listes de désinences (partie déclarative) et qui contiendrait les règles contextuelles pour la réalisation des alternances (partie procédurale)[15]. Avec ce modèle, il est aisé de générer des formes comme *olg-e,* uniquement par une simple concaténation de la désinence au radical, tout en gardant l'information qu'une alternance obligatoire n'a pas été effectuée. Cette méthode reste fidèle à l'hypothèse fondamentale du diagnostic qui essaye de procéder de la même manière que l'apprenant pour recréer la forme erronée – le modèle computationnel est pratiquement identique à la modélisation du système dans les approches structuralistes classiques ainsi qu'à la présentation pédagogique de la déclinaison dans les méthodes d'enseignement.

4.2 Description formelle des erreurs

Un autre problème lié à la norme et la normalisation était la description des erreurs observées dans les tâches de déclinaison. Des projets de taxonomie ont été entrepris dès les premières analyses des erreurs dans les recherches en acquisition d'une langue étrangère et ils se sont poursuivis dans les analyses de corpus électroniques des productions d'apprenants, voir par ex. Granger et al. (2001). Au sein de CETLEF.fr, nous avons abordé ce problème en le divisant en deux questions : (1) comment peut-on décrire une PE avec les informations morphologiques disponibles ; (2) à quelle étape, dans le processus de la production d'une PE, une erreur survient-elle ? Autrement dit, il est d'abord nécessaire d'identifier et de décrire les propriétés linguistiques observables des PE avant de chercher leur motivation ou cause probable.

Bien que les formes nominales casuelles soient principalement l'expression des fonctions syntaxiques, la description des PE a été effectuée uniquement sur la base des informations disponibles sur le niveau graphique et morphologique, intégrés dans l'annotation de la FR. Comme nous l'avons montré plus haut, le diagnostic d'une PE est fondé par la comparaison des valeurs des attributs qui diffèrent dans les structures de traits de la FR et de la FH égale à PE (voir figure 4). Ainsi, la PE *olge* peut être désignée comme une erreur d'alternance. S'il y avait d'autres attributs atteints par l'erreur, cette PE serait désignée d'après ces attributs (par exemple une erreur de cas, une erreur de cas et d'alternance, une erreur de cas, d'alternance et de nombre et ainsi de suite). Comme dans tous les domaines du TAL dont l'objectif est l'analyse des formes et structures linguistiques, le problème majeur réside dans leur ambiguïté – un élément ambigu (homonyme) peut être analysé de plusieurs manières. La situation n'est pas différente pour l'analyse morphologique des PE où le nombre de différentes interprétations peut être assez élevé à cause de l'homonymie des désinences casuelles. Pour prendre un exemple, la forme *olge* a été analysée comme une erreur d'alternance,

[15] Pour information, ce modèle, encodé en XML et réutilisable pour d'autres applications, contient 125 types de déclinaison (construits à partir des types traditionnels mais prenant en compte également les diverses variantes) et 32 règles pour les alternances consonantiques et vocaliques, enrichies par des listes d'exceptions. Actuellement, ce modèle n'est pas disponible publiquement.

mais elle pourrait être également analysée comme une erreur de cas et d'alternance, car au sein du même paradigme, la désinence *-e* exprime également le locatif. Il est évident que la première interprétation est plus probable que la seconde d'où vient la solution, adoptée actuellement pour le diagnostic, de filtrer les interprétations possibles en fonction du nombre d'attributs atteints par l'erreur – moins ce nombre est élevé, plus l'interprétation est probable.

Cependant, ce principe de désambiguïsation ne peut pas être utilisé pour une PE avec deux ou plusieurs interprétations possibles ayant le même nombre d'attributs atteints par l'erreur. Pour donner un exemple, une PE peut être interprétée soit comme erreur de cas, soit comme erreur de nombre. Dans son implémentation actuelle, le diagnostic contient un score pour chaque trait exprimant sa pertinence pour le diagnostic : une erreur de cas est préférée à une erreur de nombre, une erreur de diacritique est moins pertinente qu'une erreur de cas ou de nombre, etc. Ce classement est fondé uniquement sur des intuitions, car il est difficile de décider à quelle étape, lors de la production vue comme un processus dans la perspective psycholinguistique, l'apprenant a commis une opération erronée.

Ces questionnements appartiennent déjà à la résolution du second problème mentionné plus haut et ils ne concernent pas seulement le filtrage des interprétations possibles. Une erreur de cas, identifiée comme telle au niveau morphologique, signifie-t-elle que l'apprenant ne sait pas quel est le cas imposé par le verbe qui régit le substantif, ou est-ce que c'est un choix erroné de désinence au sein d'un paradigme pour exprimer le cas correspondant, étant donné que l'apprenant connaît la valence du verbe ? Pour la rétroaction, serait-il plus approprié dans ce cas d'attirer son attention sur le cadre verbal ou sur le paradigme de déclinaison ? Cette question pourrait être résolue facilement par un tuteur humain qui peut prendre en compte le contexte syntaxique de la PE ou les interférences de la langue maternelle de l'apprenant, mais elle est insoluble avec le diagnostic automatique disposant uniquement des informations d'ordre morphologique.

Une possibilité de rendre le diagnostic plus puissant serait de ne pas annoter uniquement la FR mais d'introduire également les informations d'ordre syntaxique en commençant par annoter la valence de l'élément qui la régit. Ceci nous permettrait, par exemple, de préférer une erreur de cas dans une situation, où il serait probable qu'un apprenant confonde les différentes structures d'arguments d'un verbe polysémique. Sous condition de disposer de l'information équivalente dans la langue maternelle de l'apprenant, il serait également possible d'identifier automatiquement les interférences interlinguistiques. En profitant des ressources linguistiques existantes, ce service pourrait être rendu par exemple par l'intégration du dictionnaire électronique de la valence verbale tchèque VALLEX, voir Lopatková et al. (2008) et Žabokrtský et Lopatková (2007)[16]. Ce lexique est fondé sur le PDT et il a été élaboré en accord avec les prémices théoriques de la description fonctionnelle générative, voir Sgall et al. (1986).

5. Conclusion

Par définition, ELAO est un domaine dont le développement est étroitement lié au progrès et à l'expansion des nouvelles technologies informatiques. A côté des aspects didactiques et pédagogiques liés à l'enseignement des langues à l'aide de l'ordinateur, l'intégration du TAL pour

[16] Le dictionnaire est disponible publiquement sur http://ufal.mff.cuni.cz/vallex/2.0/

l'analyse des productions des apprenants introduit un lien nouveau entre ELAO et les sciences du langage. La prise en compte des résultats de la recherche en acquisition d'une langue étrangère peut servir à l'adaptation des outils d'analyse automatique, développés a priori pour le traitement de la langue standard. En sens inverse, la nécessité de formalisation, liée à l'implémentation informatique, permet de prendre un regard différent sur des problèmes étudiés depuis des décennies. Sur l'exemple du diagnostic des erreurs de déclinaison en tchèque, nous espérons d'avoir montré à quel point il est utile de rester ouvert vis à vis des disciplines et des approches différentes lors de la résolution des problèmes concrets et de rester réaliste quant aux objectifs visés. L'existence des ressources linguistiques normalisées est importante pour pouvoir bâtir sur le travail déjà accompli, mais il est également nécessaire de considérer les besoins spécifiques des applications qui peuvent mener à des nouvelles solutions.

Bibliographie

ANTONIADIS, G. et al. (2005). Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. *ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication*, 8:65–79.

BRUILLARD, E. (1997). *Les machines à enseigner*. Hermès, Paris.

ČERMÁK, F. (1997). Czech national corpus: A case in many contexts. *International Journal of Corpus Linguistics*, 2(2):181–197.

DEMAZIÈRE, F. (2007). Didactique des langues et TIC : les aides à l'apprentissage. *ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication*, 10.

DESMET, P. (2006). L'enseignement/apprentissage des langues à l'ère du numérique : tendances récentes et défis. *Revue française de linguistique appliquée*, XI(1):119–138.

DODIGOVIC, M. (2005). *Artificial Intelligence in Second Language Learning*. Multilingual Matters, Clevedon.

GAONAC'H, D. (1991). *Théories d'apprentissage et acquisition d'une langue étrangère*. Hatier / Didier, Paris.

GRANGER, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Aijmer, K., éd., *Corpora and Language Teaching*. Benjamins, Amsterdam - Philadelphia, pp. 13–23.

GRANGER, S. et al. (2001). Analyse de corpus d'apprenants pour l'ELAO basé sur TAL. *Traitement automatique des langues*, 42(2):609–622.

HAJIČ, J. (2004). *Disambiguation of Rich Inflection. Computational Morphology of Czech.* Karolinum, Praha.

HAJIČ, J. (2005). Complex corpus annotation: The prague dependency treebank. In Šimková M., éd., *Insight into Slovak and Czech Corpus Linguistics*, Veda, Bratislava, pp. 54-73.

HAMEL, M.-J. & VANDEVENTER, A. (1998). Fipsgram : un analyseur syntaxique dans un contexte d'ELAO. In *Le traitement automatique du langage et les applications industrielles*. TAL + AI 98, I, Moncton, pp. 18–24.

HEIFT, T. & SCHULZE, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogue*s. Routledge, UK.

HEIFT, T. & SCHULZE, M. (2003). Error diagnosis and error correction in CALL. *CALICO*, 20(3):433–436.

JOHANNESSEN, J.-B. et al. (2002). The performance of a grammar checker with a deviant language input. In *Proceedings of the 19th international conference on Computational linguistics,* volume 2, pp 1–8.

JURAFSKY, D. and MARTIN, J. H. (2009). *Speech and Language Processing*. Pearson Education, Upper Saddle River, New Jersey.

KRAIF, O. et al. (2004). NLP tools for CALL: the simpler, the better. In *Proceedings of InSTIL / ICALL2004 – NLP and Speech Technologies in Advanced Language Learning Systems,* Venice.

KRAIF, O. and PONTON, C. (2007). Du bruit, du silence et des ambiguïtés : que faire du TAL pour l'apprentissage des langues ? In *Actes de TALN 2007*, Toulouse, France.

LEVY, M. (1997). *Computer-Assisted Language Learning: Context and Conceptualization*. Clarendon Press, Oxford.

LOPATKOVÁ, M. et al. (2008). *Valenční slovník českých sloves.* Nakladatelství Karolinum, Prague.

L'HAIRE, S. and VANDEVENTER-FALTIN, A. (2003). Diagnostic d'erreurs dans le projet FreeText. *ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication*, 6(2):21–37.

MEURERS, D. (2009). On the Automatic Analysis of Learner Language. *CALICO Journal,* 26(3):469–473.

NERBONNE, J. (2003). Natural language processing in computer-assisted language learning. In Mitkov, R., éd., *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 670–698.

PIRELLI, V. (2007). Psycho-computational issues in morphology learning and processing : An ouverture. *Lingue e Linguagguo*, VI(2):131–138.

PORQUIER, R. (1977). L'analyse des erreurs. Problèmes et perspectives. *Étude de linguistique appliquée*, 25:23–43.

RIMROTT, A. (2003). A spell checking algorithm for treating predictable verb inflection mistakes made by non-native writers of german. In *LING 807 – Computational Linguistics at Simon Fraser University*, Burnaby, Canada.

SCHNEIDER, D. and MCCOY, K. F. (1998). Recognizing syntactic errors in the writing of second language learners. In *Proceedings of COLING-ACL*, pp. 1198–1204.

SCHULZE, M. (2008). Modeling SLA processes using NLP. In Chapelle, C. et al., éds., *Towards adaptive CALL: Natural language processing for diagnostic language assessment*. Iowa State University, pp. 149–166.

SEDLÁČEK, R. and SMRŽ, P. (2001). A new czech morphological analyser ajka. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, Springer-Verlag, London, pp. 100–107.

SGALL, P., HAJIČOVÁ, E., and PANEVOVÁ, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Praha – Amsterdam.

SPOUSTOVÁ, D. (2008). Combining statistical and rule-based approaches to morphological tagging of czech texts. *The Prague Bulletin of Mathematical Linguistics,* (89):23–40.

ŠMILAUER, I. (2008). *Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison*. Thèse PhD, INALCO / Faculté des lettres de l'Université Charles, Paris / Prague.

ŽABOKRTSKÝ, Z. and LOPATKOVÁ, M. (2007). Valency Information in VALLEX 2.0. Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.