Acquisition du tchèque par les francophones: Analyse automatique des erreurs de déclinaison

Ivan Šmilauer

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 90 DECEMBER 2008 33–56

Acquisition du tchèque par les francophones: Analyse automatique des erreurs de déclinaison

Ivan Šmilauer

Abstract

This paper is a summary of our PhD thesis (Šmilauer, 2008) that presents the concept and the implementation of a platform of computer-assisted language learning, featuring on-line fill-in-the-blank exercises with feedback on errors in Czech declension (www.cetlef.fr). Morphological annotation of required forms enables a didactic presentation of the morphological system on the learning platform, as well as the implementation of a procedure of automatic error diagnosis that is carried out by the comparison of an erroneous production with hypothetical forms generated from the stem of the required form. The device can be used as a source of data for a research into second langauge acquisition.

1. Introduction

La langue tchèque, avec sa flexion très riche, offre un matériau intéressant du point de vue de l'acquisition de la morphologie par des étudiants étrangers.

Dans notre thèse, nous nous sommes concentrés sur les erreurs commises par les apprenants francophones dans un cadre expérimental restreint : celui d'exercices de déclinaison dans lesquels il faut décliner le lemme d'une forme (substantif, adjectif, pronom, numéral) au sein d'une phrase.

Une erreur de déclinaison se manifeste par un écart entre une ou plusieurs formes requises, dans un certain contexte syntaxique, et la forme erronée produite par l'apprenant. Nous avons établi l'hypothèse qu'une telle forme peut être générée automatiquement à partir du lemme de la forme requise, à l'aide des moyens formels de la déclinaison (choix d'une désinence, réalisation des alternances), employés d'une manière incorrecte. En nous basant sur cette hypothèse, nous avons proposé un module de diagnostic automatique des erreurs dont l'objectif est de générer un message de retour spécifiant le type de l'erreur au niveau morphologique.

Ce diagnostic, possible grâce à l'annotation morphologique des formes requises, a été implémenté sur une plateforme d'enseignement de langue assisté par ordinateur qui représente la

Please cite this article as: Ivan Šmilauer, Acquisition du tchèque par les francophones: Analyse automatique des erreurs de déclinaison. The Prague Bulletin of Mathematical Linguistics No. 90, 2008, 33–56.

partie appliquée de notre thèse. Cette plateforme, nommée CETLEF[1], est une application Web dynamique (disponible librement sur www.cetlef.fr) contenant une base de données relationnelle gérée par MySQL et une interface XHTML avec des éléments dynamiques en Javascript. Les procédures automatiques sont implémentées en langage PHP. CETLEF contient une plateforme auteur qui sert pour la création des exercices, et une plateforme apprenant qui est destinée aux étudiants.Pendant l'inscription sur cette plateforme, les apprenants fournissent des informations qui peuvent aider pendant l'interprétation de leur productions (âge, durée de l'apprentissage du tchèque, autres langues maîtrisées, etc.).

2. Motivation de CETLEF

Bien que l'enseignement pratique du tchèque langue étrangère (TLE) soit d'une tradition relativement riche, voir (Hrdlička, 2002), ce n'est qu'à partir des années 1980 que commencent à apparaître des travaux préliminaires, incitant à la constitution d'un champ de recherche autonome dont l'objet serait une méthodologie spécifique pour l'enseignement du TLE. La présentation didactique de la déclinaison est un des problèmes principaux dans l'enseignement, voir par exemple (Poldauf and Špruňk, 1968) ou (Nekula, 2007).

2.1. CETLEF comme source de données pour l'analyse des erreurs

En adoptant l'analyse des erreurs comme moyen privilégié pour étudier l'acquisition d'une langue étrangère, voir par exemple (Porquier, 1977), (Besse and Porquier, 1991), (Gaonac'h, 1991), le premier de nos objectifs a été de concevoir un outil qui puisse servir comme source de données pour l'étude des erreurs dans la déclinaison.

2.1.1. Productions libres

L'avantage des *productions libres* est l'authenticité des données qui reflètent l'emploi effectif de la langue. L'inconvénient principal est une collecte de données coûteuse en temps et effort. Les productions libres sont également affectées par la volonté d'utiliser des structures et un vocabulaire que l'apprenant estime maîtriser suffisamment bien pour pouvoir s'en servir dans la communication. On parle dans ce cas des «stratégies d'évitement», voir par exemple (Porquier, 1977), (Bautier-Castaing, 1977).

Afin de disposer de plus de données pour l'analyse, les collectes de corpus électroniques de productions d'apprenants commencent à émerger depuis une quinzaine d'années, voir (Granger, Hung, and Petch-Tyson, 2002), (Pravec, 2002), (Tono, 2003).

2.1.2. Productions issues des exercices

Les *données sollicitées* dans un cadre expérimental, permettant de mieux contrôler les facteurs situationnels, doivent nécessairement contenir les phénomènes spécifiques qui ont été

[1] Acronyme de *Connaître / Comprendre / Corriger les Erreurs en Tchèque Langue Étrangère pour les Francophones.*

établis comme objet de l'investigation. Néanmoins, l'authenticité de ces données peut être contestée, ainsi que leur ambition de refléter l'état réel de la compétence de l'apprenant. Par rapport à un corpus de productions libres, le recueil de données produites dans des exercices ciblés sur une compétence spécifique peut apporter plus rapidement des données pertinentes. Les informations sur l'emploi d'une certaine structure, obtenues à l'aide des exercices, seraient beaucoup plus éparses dans un corpus de productions libres et le nombre de leurs occurrences serait proportionnel à sa taille.

2.1.3. Arguments pour les exercices

CETLEF permet de collecter les données au sein d'exercices grammaticaux contenant des tâches de déclinaison : les formes requises dans de telles tâches peuvent être facilement accompagnées par une annotation morphologique (la catégorie lexicale, les catégories morphologiques, le type paradigmatique et l'indication d'une éventuelle alternance), ajoutée manuellement ou avec des méthodes semi-automatiques lors de la création des exercices. Le stockage des productions dans une base de données relationnelle permet leur exploitation efficace. De plus, la plateforme proposant des exercices peut intégrer des fonctionnalités supplémentaires à visée didactique.

2.2. CETLEF comme un outil d'enseignement de langue assisté par ordinateur

L'enseignement ou l'apprentissage des langues assisté par ordinateur (ELAO ou ALAO, CALL pour Computer Assisted Language Learning) est un domaine pluridisciplinaire dont l'objet est l'intégration d'outils informatiques dans l'enseignement des langues, pour des revues synthétiques sur la discipline, voir par exemple (Levy, 1997), (Nerbonne, Jager, and van Essen, 1998), (Cameron, 1999), (Hanson-Smith, 2003). Ces outils sont considérés plutôt comme un complément de l'enseignement traditionnel qu'une alternative à celui-ci, voir (Bertin, 2001).

D'après (Karttunen, 1986), (Zock, 1996), (Nerbonne, 2003) et d'autres, l'enseignement des langues assisté par ordinateur est un domaine idéal pour la vérification des fonctionnalités des techniques de TAL, car la tâche d'assister un apprenant dans son apprentissage implique virtuellement tous les objectifs visés par cette discipline. L'intégration de messages de diagnostic des erreurs, dans les outils d'enseignement assistés par ordinateur, est considéré comme un point positif au niveau didactique, voir par exemple (Heift and Schulze, 2003), (L'haire and Vandeventer-Faltin, 2003), (Heift and Schulze, 2007).

Des méthodes de correction et de diagnostic des erreurs peuvent être appliquées soit sur des productions libres, soit sur des productions provenant de tâches fermées comme dans les exercices grammaticaux. Pour le traitement des productions libres, différentes techniques sont expérimentées pour adapter les correcteurs orthographiques et grammaticaux, destinés à l'usage universel, afin qu'ils puissent prendre en compte les spécificités des textes produits par des apprenants étrangers.

Par rapport à l'imperfection actuelle des outils disponibles pour la correction des productions libres, (Holland and Kaplan, 1995), (Kraif et al., 2004), (Tschichold, 2006) estiment néces-

saire l'adoption d'une approche «pédagogiquement responsable», favorisant l'emploi de techniques de base qui sont suffisamment bien maîtrisées pour réduire le bruit ou le silence à la sortie du traitement. Ces imperfections, qui peuvent être acceptables pour certaines applications dans leur usage «non pédagogique», se révèlent particulièrement perturbantes pour un apprenant au sein d'un didacticiel.

Dans la perspective de l'emploi des techniques de TAL pour la correction des erreurs dans un outil ELAO, nous estimons qu'un diagnostic des erreurs, issues des exercices grammaticaux à trous, peut être effectué par des procédés relativement simples et fiables, basés sur la génération morphologique.

Fig. *1. Page d'entrée de la plateforme apprenant de CETLEF*

3. Annotation morphologique

Dans le cadre de CETLEF, nous avons élaboré un modèle de la déclinaison du tchèque répondant aux objectifs suivants :

1. Annoter les productions des apprenants par des métadonnées linguistiques qui seraient utiles pour leur analyse. Comme dans le cas des corpus annotés, il s'agit de **faciliter la**

recherche de l'information à l'aide des étiquettes linguistiques.

2. Servir pour le **diagnostic automatique des erreurs**. L'indication des valeurs des catégories grammaticales, du type paradigmatique et de l'alternance est une information cruciale pour l'interprétation automatique d'une erreur de déclinaison.
3. Être affiché, dans un format adapté pour une **présentation didactique**, sur la plateforme apprenant en tant qu'assistance dans l'apprentissage.

3.1. Définition des types paradigmatiques

Le classement des types paradigmatiques a été fondé sur la tradition grammaticale tchèque (14 types de déclinaison de bases), nous définissons cependant une classification détaillée des différents sous-types et des exceptions. Contrairement aux autres annotations morphologiques utilisées pour le traitement automatique du tchèque, voir (Hajič, 2004) et (Osolsobě, 1996), cette classification est établie uniquement par les différences dans les ensembles de désinences qui s'attachent au radical du lexème pour créer une forme déclinée. La réalisation des alternances vocaliques ou consonantiques n'est pas prise en compte pour la définition des types paradigmatiques ce qui permet de diminuer leur nombre.

3.1.1. Représentation d'un paradigme

Un exemple de la représentation complète du paradigme de désinences casuelles d'un certain sous-type est présenté sur la table 1. Cette table représente le paradigme de désinences casuelles du sous-type *hoch*, défini par rapport au sous-type modèle *pán* du type *pán*. Il se distingue par le remplacement (valeur de l'attribut ET - écart du type) de la désinence *-e* par *-u* dans le vocatif singulier (*pan-e* × *hoch-u*), par le remplacement de la désinence *-ech* par *-ích* dans le locatif pluriel et par l'ajout d'une variante de registre *-ách* pour le même cas. Les désinences qui sont des variantes fonctionnelles ou des variantes de registre sont marquées par la valeur correspondante de l'attribut var (variante).

3.2. Alternances vocaliques et consonantiques

Afin de pouvoir annoter la réalisation des alternances vocaliques ou consonantiques, et de déterminer les règles de leur réalisation, nous avons entrepris une étude basée sur l'ensemble des 50 000 lexèmes du tchèque les plus fréquents dans (Čermák and Křen, 2004). Les différentes configurations dans lesquelles une alternance peut être effectuée ont été examinées afin de définir les règles de leur réalisation et des listes d'exceptions.

3.3. Annotation des tâches dans les exercices

Avec ce répertoire d'étiquettes morphologiques, formatées dans des fichiers XML *pdgm.xml* pour les types paradigmatiques et *alt.xml* pour les alterances, la forme requise dans une tâche est annotée manuellement par l'auteur des exercices. Vu le nombre de tâches qui sont proposées

cas	num	gen	var	ET	term
nom	sg	m			#
gen	sg	m			a
dat	sg	m	fnct		u
dat	sg	m	fnct		ovi
acc	sg	m			a
voc	sg	m		R	u
loc	sg	m	fnct		u
loc	sg	m	fnct		ovi
inst	sg	m			em

cas	num	gen	var	ET	term
nom	pl	m	fnct		i
nom	pl	m	fnct		ové
gen	pl	m			ů
dat	pl	m			ům
dat	pl	m	reg		um
acc	pl	m			y
voc	pl	m	fnct		i
voc	pl	m	fnct		ové
loc	pl	m		R	ích
loc	pl	m		A	ách
inst	pl	m			y
inst	pl	m	reg		ama

Tab. *1. Représentation du paradigme de désinences casuelles d'un sous-type paradigmatique (Valeurs de l'attribut* `var` *:* `fnct` *- variante fonctionnelle,* `reg` *- variante de registre)*

dans le cadre d'une enquête contenant des exercices, une annotation manuelle ne représente pas un travail inabordable.

L'annotation d'une forme requise, telles qu'elle est stockée dans la base de données, contient les informations suivantes :

requis	lemme	tagLex	tagMorph	pdgm	cas	num	gen	alt
hub	houba	subst	N	zn	gen	pl	f	ou > u

Tab. *2. Exemple de l'annotation d'une forme requise*

Explication des noms des attributs : `requis` : *la forme requise*, `lemme` : *lemme*, `tagLex` : *catégorie lexicale*, `tagMorph` : *type morphologique*, `pdgm` : *type paradigmatique*, `cas` : *cas*, `num` : *nombre*, `gen` : *genre*, `alt` : *alternance*.

3.4. Exploitation didactique de l'annotation

Une des fonctions principales du modèle de la déclinaison est l'affichage des informations d'ordre didactique à l'apprenant. L'annotation des formes requises sont affichées pendant la correction de l'exercice. Pour l'exemple, voir la figure 2 avec l'affichage de l'annotation de la forme requise, dont l'annotation a été présentée dans la table ci-dessus (tab. 2). Il s'agit de la correction d'un exercice effectué.

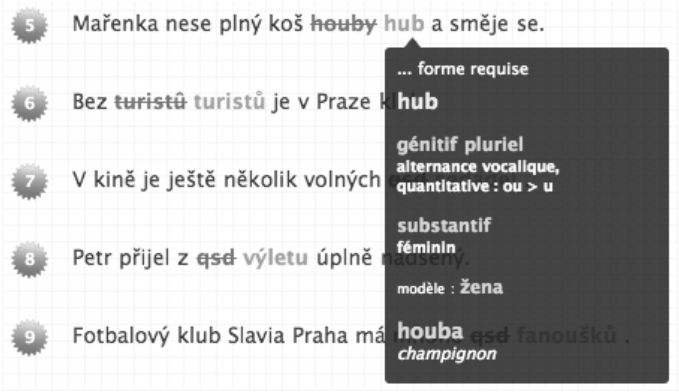

Fig. *2. Annotation d'une forme requise sur la plateforme apprenant*

Les éléments linguistiques que l'apprenant rencontre dans les exercices, remplissent sa base de données personnelle qui peut être consultée à tout moment et qui peut servir comme outil d'apprentissage. L'augmentation du volume de cette base peut être un facteur motivant pour l'apprenant. Les éléments qui nourrissent cette «base de connaissances» sont classées en quatre sections : le lexique contenu dans les exercices terminés ; les paradigmes des formes à décliner ; les alternances à effectuer dans les formes à décliner ; le cas et le nombre des formes à décliner.

3.4.1. Lexique

La section *Lexique* contient une liste alphabétique des mots tchèques et de leur traduction française qui apparaissent dans les exercices soit en tant que formes requises, soit comme appartenant au contexte gauche ou droit d'une phrase donnée. Le lien hypertexte derrière chaque entrée permet d'afficher la tâche dans laquelle cette entrée est apparue.

3.4.2. Paradigmes

Dans la section *Paradigmes*, l'apprenant peut choisir l'un des types paradigmatiques pour afficher la liste des terminaisons (voir fig. 3) et les mots rencontrés dans les exercices qui appartiennent à ce paradigme. Seules les terminaisons du modèle de type paradigmatique sont présentées, les variantes du registre ne sont pas affichées. Les différences dans la déclinaison des mots dans le lexique par rapport à leur type modèle sont mises en évidence (voir fig. 4).

3.4.3. Alternances

La section *Alternances* permet de visualiser les différentes occurrences des alternances rencontrées dans les tâches des exercices. Pour afficher les exemples de la réalisation d'une al-

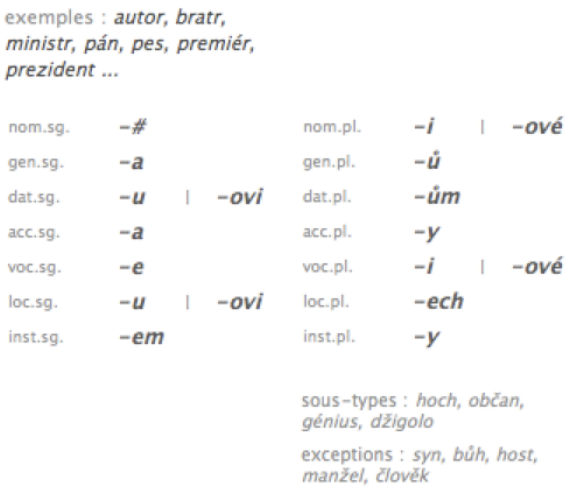

Fig. *3. Présentation d'un type paradigmatique*

Fig. *4. Affichage des différences dans la déclinaison d'un sous-type*

ternance dans les exercices, l'apprenant est invité d'abord à choisir un type d'alternance (par exemple vocalique quantitative, palatalisation A, mouillure etc.), puis une alternance particulière (par exemple *k* > *c*). Au passage de la souris sur le couple *lemme – forme alternée*, il est possible de visualiser une vignette portant des informations sur le mot alterné et l'alternance elle même (voir fig. 5).

Fig. *5. Vignette accompagnant une alternance*

3.4.4. Cas

La section *Cas* permet de consulter les formes requises dans les exercices en fonction du nombre grammatical et du cas (voir fig. 6).

Fig. *6. Consultation des formes requises par nombre et par cas*

4. Diagnostic des erreurs

Dans le diagnostic automatique des erreurs, une production erronée est considérée comme une combinaison inappropriée du radical de la forme requise et d'une désinence casuelle. Grâce

à l'annotation d'une forme requise, il est possible de générer automatiquement différentes formes hypothétiques qui pourraient être produites par un apprenant et ces formes hypothétiques sont comparées à la production erronée. S'il y a une correspondance, la production erronée est interprétée par les différentes propriétés de la forme hypothétique correspondante. Cette approche est basée sur les hypothèses présentées ci-dessous.

4.1. Hypothèses sur les erreurs possibles

Les hypothèses sur les erreurs possibles, commises dans les exercices de déclinaison par des apprenants francophones, peuvent être établies sur la base de la comparaison du système nominal tchèque et français. La déclinaison du tchèque présente pour un apprenant français, ou pour tout autre apprenant dont la langue maternelle ne dispose pas de la déclinaison, une sorte d'**idiosyncrasie** par rapport au système de sa langue maternelle. La variation des formes fléchies rajoute de la complexité dans la production langagière et peut être naturellement une source d'erreurs.

Du point de vue de l'activité de production langagière de l'apprenant, il est possible de distinguer plusieurs étapes, qui sont nécessaires pour produire une forme casuelle correcte au sein d'une tâche de déclinaison et qui peuvent mener à l'erreur : (1) le choix des valeurs de la catégorie du cas, du nombre et du genre ; (2) le classement du lexème dans le paradigme approprié et le choix de la désinence correspondante aux valeurs des catégories grammaticales ; (3) la réalisation, si cela est nécessaire, des alternances vocaliques ou consonantiques.

4.1.1. Choix des valeurs des catégories grammaticales

L'étape (1) est effectuée en fonction des critères purement syntaxiques pour l'attribution du cas. Les valeurs du genre et du nombre sont attribuées en fonction de l'accord entre la forme requise et son régisseur où en fonction des critères d'ordre sémantiques, relatifs au contenu cognitif exprimé par la phrase. Un dysfonctionnement dans cette opération serait reflété dans la production par le choix de désinences exprimant les valeurs inappropriées des catégories respectives.

4.1.2. Attribution du paradigme casuel au lexème

L'attribution d'un certain type paradigmatique au lemme de la forme requise pendant l'étape (2) délimite le répertoire des désinences permettant d'exprimer les valeurs des catégories grammaticales choisies à l'étape précédente. Nous pouvons supposer l'existence de formes produites par les apprenants qui peuvent être correctes, en ce qui concerne les valeurs des catégories grammaticales liées à la désinence choisie pour la génération d'une certaine forme, mais qui ne sont pas appropriées pour exprimer les significations grammaticales au sein du paradigme propre à la forme requise. Par exemple, le génitif pluriel *turist* du lexème *turista* n'est pas généré d'après le paradigme approprié (type *předseda*), demandant la désinence *-ů* dans le génitif pluriel, mais d'après le type *žena* qui emploie effectivement la désinence zéro -# pour exprimer le cas et le nombre correspondants.

4.1.3. Réalisation des alternances

L'étape (3) représente une opération sur la forme composée du radical et la désinence. La réalisation des alternances est conditionnée par des facteurs phonologiques, morphologiques et lexicaux et la maîtrise des règles de leur réalisation est nécessaire pour la création adéquate des formes casuelles. Nous pouvons supposer, qu'un apprenant fera des erreurs dans les alternances, comme par exemple *houb* au lieu de *hub* dans le génitif pluriel du substantif *houba* (*champignon*), etc.

4.2. Définition des types d'erreurs

Les différents types d'erreurs sont définis sur la base des propriétés morphologiques des formes hypothétiques, générées à partir du radical de la production requise et des désinences employées pour la génération des formes casuelles du tchèque.

4.2.1. Définition préliminaires

1. Soit un alphabet L, ensemble fini de caractères ; soit un langage $L^\star$, ensemble infini de toutes les chaînes possibles sur l'alphabet L ;
2. Soit un alphabet $L_{CZ} \subset L$, ensemble fini contenant tous les caractères du tchèque à part l'espace, les chiffres et les signes de ponctuation, et qui est une union des ensembles de caractères minuscules, majuscules, minuscules diacritées et majuscules diacrités ; soit un langage $L_{CZ}{}^\star \subset L^\star$, ensemble infini de toutes les chaînes possibles sur l'alphabet L_{CZ} ;
3. Soit une fonction Min, opération de minusculisation qui attribue à chaque mot $m \in L_{CZ}{}^\star$ sa forme correspondante uniquement en minuscules ; soit une fonction Dia, opération d'enlèvement du diacritique qui attribue à chaque mot $m \in L_{CZ}{}^\star$ sa forme correspondante en caractères sans diacritique ; soit une fonction St, opération de standardisation telle que $St(m) = Dia(Min(m))$;
4. Soit un langage $N \subset L_{CZ}{}^\star$, ensemble fini de toutes les formes lexicales des mots tchèques appartenant aux types morphologiques nominal, adjectival, adjectival mixte, pronominal ou numéral : N = {*abatyše, ..., mládě, mláděte, ..., žížalami*}.
5. Soit un ensemble $R \subset L_{CZ}{}^\star$, ensemble fini de tous les radicaux extraits par l'enlèvement de la désinence casuelle de la forme du lemme des mots tchèques appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal ou numéral ; soit un ensemble $D \subset L_{CZ}{}^\star$, ensemble fini de toutes les désinences des types paradigmatiques des mots tchèques appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal.
6. Soit un langage $H \subset L_{CZ}{}^\star$, ensemble fini de toutes les formes lexicales hypothétiques des mots tchèques h appartenant au type morphologique nominal, adjectival, adjectival mixte, pronominal ou numéral ; ainsi que leurs formes minusculisées $Min(h)$, sans diacritique $Dia(h)$ et standardisées $St(h)$. Ces formes sont le résultat de la concaténation

des couples contenus dans le produit des ensembles $R \times D$, avec ou sans la réalisation de l'alternance sur la chaîne résultante

4.2.2. Définition de la forme requise et de la production erronée

1. La **forme requise** r dans une tâche x est un mot tel que $r \in N$. Chaque r est caractérisée par son annotation morphologique.
2. La **production erronée** p dans une tâche x est un mot tel que $p \in L^{\star}$ et $p \neq r$.
3. Une production erronée p **peut être interprétée morphologiquement** si p correspond à une des formes lexicales hypothétiques $h \in H$, générées à partir du radical de la forme requise r.
4. Une production erronée p **ne peut pas être interprétée morphologiquement** si p ne correspond à aucune des formes hypothétiques $h \in H$, générées à partir du radical de la forme requise r.

4.3. Interprétation morphologique

Chaque forme requise r dans le cadre d'une tâche x est caractérisée par son annotation morphologique. Pour les besoins de la description formelle des erreurs, cette annotation peut être représentée à l'aide d'une **structure de traits**. Pour une forme requise r, la structure de traits est la suivante :

$$\begin{bmatrix} cas : cas \\ num : nombre \\ gen : genre \\ alt : identifiant\,de\,l'alternace \\ pdgm : soustype\,paradigmatique \\ tagMorph : type\,morphologique \end{bmatrix}$$

Les différentes valeurs des attributs dans cette structure sont instanciées en fonction des propriétés morphologiques de r inscrites dans l'annotation. Les noms des attributs sont identiques à ceux utilisés dans l'annotation morphologique sur CETLEF.

Par exemple, pour une forme requise *r = matce* qui est le datif singulier du substantif *matka*, l'instanciation de la structure de traits est la suivante :

$$\begin{bmatrix} cas : dat \\ num : sg \\ gen : f \\ alt : k > c \\ pdgm : zn\_Re \\ tagMorph : N \end{bmatrix}$$

L'ensemble des formes hypothétiques h, générées à partir du radical *matka*, est créé par toutes les combinaisons possibles du radical *matk* avec toutes les désinences dans D, avec ou sans la réalisation des alternances, avec ou sans diacritique et avec toutes les possibilités dans la casse des caractères : *matka, matky, ..., matek, matk, ..., matkám, matkam, ..., matkách, matkach, ..., matkovi, matkem, ..., matkému, ...*

À chacune de ces formes h peut être assignée au moins une structure de traits. Les valeurs des attributs dans la structure de h sont déterminées uniquement sur la base de propriétés formelles de ses composants en fonction (1) des différentes valeurs des catégories morphologiques qui peuvent être exprimées par la désinence employée, (2) de la réalisation d'une alternance sur le radical ou (3) de sa forme graphique. Une structure de traits assignée à une forme h est appelée son **interprétation**.

Le nombre des interprétations des formes hypothétiques n'est pas déterminé uniquement par l'homonymie des formes casuelles existantes pour un certain lemme, mais également par toutes les combinaisons possibles du radical et des désinences appartenant aux autres paradigmes, ainsi que par les variation dans la diacritique.

4.4. Erreur d'après l'attribut atteint

En fonction des attributs qui diffèrent dans les structures de r et de p (qui sont atteints par l'erreur), les différents types d'erreurs sont représentés dans la table 3.

attribut	type d'erreur
cas	erreur de cas
num	erreur de nombre
gen	erreur de genre
alt	erreur d'alternance
pdgm	erreur de type paradigmatique
pdgm	erreur de sous-type paradigmatique
tagMorph	erreur de type morphologique
dia	erreur de diacritique
casse	erreur de casse

Tab. *3. Erreurs d'après l'attribut atteint*

Ces erreurs peuvent se combiner librement entre elles en fonction des attributs atteints par l'erreur dans une interprétation donnée. Il peut exister par exemple une erreur de cas et de nombre, une erreur de cas et d'alternance, une erreur de nombre et de graphie, etc. Dans ces appellations, chaque attribut qui contient une valeur différente par rapport à la forme requise doit être spécifié.

4.5. Erreur par rapport au paradigme de la forme requise

Sur la base des observations des erreurs authentiques produites par les apprenants dans les exercices de déclinaison, nous avons établi quatre groupes dans lesquels l'ensemble des interprétations des formes hypothétiques h, employées pour la recherche d'une correspondance avec une production erronée p pour une forme requise r, établi par rapport au paradigme de la forme requise : erreur locale, erreur verticale, erreur horizontale interne, erreur horizontale externe.

4.5.1. Erreur locale

La désinence appartient au sous-type paradigmatique de la forme requise r avec la même valeur de cas, de genre et de nombre. Dans l'exemple (1), la production erronée *Olge* est une erreur locale d'alternance.

(1) *Petr vzal* ***Olge*** *všechny peníze.*
Olze dat.sg.f.
Pierre a pris à Olga tout argent
'Pierre a pris à Olga tout l'argent'

$$\begin{array}{c} Olze \\ \left[\begin{array}{l} cas : dat \\ num : sg \\ gen : f \\ \textbf{alt : g > z} \\ pdgm : zn\_Re \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array} \neq \begin{array}{c} Olge \\ \left[\begin{array}{l} cas : dat \\ num : sg \\ gen : f \\ \textbf{alt : sans} \\ pdgm : zn\_Re \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array}$$

4.5.2. Erreur verticale

La désinence appartient au sous-type paradigmatique de la forme requise r avec la valeur de cas autre que celle de la forme requise. Dans l'exemple (2), la production erronée *průvodci* est une erreur verticale de cas.

(2) *Výklad našeho* ***průvodci*** *byl velice zajímavý.*
průvodce gen.sg.m.
Exposé notre guide était très intéressant
'L'exposé de notre guide a été très intéressant'

průvodce		*průvodci*
cas : gen	≠	**cas : dat | loc**
num : *sg*		*num* : *sg*
gen : *m*		*gen* : *m*
alt : *sans*		*alt* : *sans*
pdgm : *sc*		*pdgm* : *sc*
tagMorph : *N*		*tagMorph* : *N*
dia : 1		*dia* : 1
casse : 1		*casse* : 1

4.5.3. Erreur horizontale interne

La désinence appartient aux autres sous-types dans le type paradigmatique de la forme requise r avec les mêmes valeurs de cas, de nombre et de genre. Dans l'exemple (3), la production erronée *výleta* est une erreur horizontale interne de sous-type paradigmatique.

(3) *Petr přijel z* ***výleta** *v Bretani.*
výletu gen.sg.i.
Petr est rentré de voyage en Bretagne

'Pierre est rentré du voyage en Bretagne'

výletu		*výleta*
cas : *gen*	≠	*cas* : *gen*
num : *sg*		*num* : *sg*
gen : *i*		*gen* : *i*
alt : *sans*		*alt* : *sans*
pdgm : hd		**pdgm : hd_l**
tagMorph : *N*		*tagMorph* : *N*
dia : 1		*dia* : 1
casse : 1		*casse* : 1

4.5.4. Erreur horizontale externe

La désinence appartient aux autres types paradigmatiques dans le cadre du même type morphologique avec les mêmes valeurs de cas et de nombre et qui ont la même désinence dans le nominatif singulier comme le lemme de r. Dans l'exemple (4) la production erronée *sole* est une erreur horizontale externe de type paradigmatique (le choix de la désinence *-e* du type *píseň* au lieu de la désinence *-i* du type *kost*).

(4) *Maso bez* ***sole** *není většinou příliš chutné*
soli gen.sg.f
viande sans sel n'est pas d'habitude très appétissant

'La viande sans sel n'est pas d'habitude très appétissante'

$$\begin{array}{ccc} soli & \neq & sole \\ \left[\begin{array}{l} cas : gen \\ num : sg \\ gen : f \\ alt : \mathring{u} > o \\ \textbf{pdgm : kt\_n} \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] & & \left[\begin{array}{l} cas : gen \\ num : sg \\ gen : f \\ alt : \mathring{u} > o \\ \textbf{pdgm : ps\_Re} \\ tagMorph : N \\ dia : 1 \\ casse : 1 \end{array}\right] \end{array}$$

4.6. Diagnostic morphologique d'une production erronée

Le diagnostic morphologique d'une production erronée p dans une tâche x est l'ensemble de ses interprétations qui sont le plus plausibles du point de vue de l'activité langagière de l'apprenant. La plausibilité d'une interprétation peut être établie sur la base des critères morphologiques pour les erreurs locales et horizontales. Par contre, pour les erreurs verticales, où la valeur de la catégorie du cas est une variable, des facteurs syntaxiques entrent nécessairement en jeu.

Le diagnostic automatique sur CETLEF ne peut prendre en compte que les informations morphologiques et son ambition n'est que de proposer la meilleure solution dans le cadre donné. Cette solution peut être par la suite confirmée ou rejetée à l'aide d'une étude «manuelle», effectuée par un humain qui prend en compte des critères divers qui lui permettent de choisir l'interprétation la plus probable.

Dans le cadre morphologique, nous définissons donc que *la plausibilité d'une interprétation est déterminée par le nombre d'attributs atteints par l'erreur*. Moins il y a d'attributs qui diffèrent dans la structure de r et dans une certaine interprétation de p, plus cette interprétation est plausible.

4.7. Description de la procédure de diagnostic des erreurs

La procédure *Diagnostic* est employée pour diagnostiquer les productions qui ne correspondent à aucune des formes requises dans le cadre d'une tâche. Ce diagnostic est effectué par la recherche de différentes interprétations de la production erronée et par le choix de celles qui sont les plus plausibles.

Les données à l'entrée de la procédure sont : la *production erronée*, la *forme requise*, le *lemme* de la forme requise et l'*annotation* de la forme requise. À la sortie de la procédure, un message qui spécifie l'erreur est généré. Ce message d'erreur sert comme critère pour les recherches des productions dans la base de données en fonction des différents types d'erreurs et pour la génération du message de diagnostic, spécifiant la nature de l'erreur sur la plateforme apprenant.

4.7.1. Traitement non morphologique

D'abord, l'interprétation de l'erreur est effectuée à l'aide des techniques simples qui n'impliquent pas l'utilisation des données morphologiques. Le but est d'identifier, à l'aide des calculs sur les

caractères et les chaînes de caractères qui représentent la forme requise et la production erronée, une différence trop importante entre ces deux éléments (distance de Levenshtein (Levenshtein, 1966) trop grande, différence importante de longueur des deux chaînes, etc.), pour qu'il puisse y avoir une interprétation morphologique . Si cette étape n'a pas été suffisante pour déterminer le bon diagnostic, une série de tests morphologiques est commencée pour interpréter l'erreur comme une forme hypothétique générée à partir du radical de la forme requise.

4.7.2. Traitement morphologique

Les tests morphologiques utilisent les informations linguistiques dans l'annotation de la forme requise, le modèle de la déclinaison, structuré dans les fichiers *pdgm.xml* et *alt.xml*, et la procédure *AlterneRadical*, qui assure les changements du radical au contact d'une désinence susceptible de provoquer une alternance vocalique ou consonantique.

Pendant ce traitement, des formes hypothétiques sont générées à partir du radical de la forme requise. Ces formes doivent nécessairement observer les restrictions posées sur les erreurs locales, verticales, horizontales internes et horizontales externes. Chaque forme hypothétique est ensuite systématiquement comparée à la production erronée. S'il y a une correspondance, l'erreur est interprétée sur la base des propriétés morphologiques de cette forme et cette interprétation est inscrite parmi les autres possibles.

4.7.3. Filtrage des interprétations

Pendant cette étape, le message d'erreur est filtré pour réduire au minimum le nombre des interprétations possibles afin d'en retenir uniquement celles qui sont les plus plausibles. Cette réduction est effectuée en fonction du nombre d'attributs atteints par l'erreur qui détermine leur classement dans l'échelle de plausibilité pour un diagnostic.

Prenons les différentes interprétations de l'erreur dans la tâche suivante :

(5) *David nemůže jíst* ****rajče.***
rajčata acc.pl.n.
David ne peut pas manger les tomates
'David ne peut pas manger des tomates'

La comparaison de la production erronée avec les formes hypothétiques détermine qu'il peut s'agir des erreurs suivantes : (a) une erreur verticale de nombre d'après l'accusatif singulier ; (b) une erreur verticale de cas et de nombre d'après le nominatif singulier ; (c) une erreur verticale de cas et de nombre d'après le vocatif singulier ; (d) une erreur horizontale externe de type paradigmatique d'après l'accusatif pluriel du type *moře* ; (e) une erreur horizontale externe de type paradigmatique et de genre d'après l'accusatif pluriel du type *růže* ; (f) une erreur horizontale externe de type paradigmatique et de genre d'après l'accusatif pluriel du type *soudce*.

Pendant l'étape de filtrage des interprétations possibles, les interprétations retenues comme les plus probables sont les interprétations (a) et (d) : l'apprenant se trompe soit dans le nombre et met la forme du singulier au lieu de la forme du pluriel ; soit il confond le type paradigmatique

kuře avec le type *moře*, qui a le même genre. Cette décision est prise sur la base du nombre de traits morphologiques atteintes par l'erreur : il s'agit d'un seul trait pour les interprétations (a) et (d) ; et de deux traits pour les interprétation (b), (c), (e) et (f). Sur la base de ce critère, les interprétations (b), (c), (e), (f) peuvent être rejetées, car deux interprétations, classées plus haut sur l'échelle de la plausibilité, ont été trouvées.

4.7.4. Formatage du diagnostic

La dernière étape de la procédure *Diagnostic* consiste en une «traduction» des interprétations retenues dans le message filtré en langue naturelle pour qu'elles puissent être publiées sur la plateforme apprenant afin de servir comme une explications des production erronées. Cette procédure est basée sur un principe simple de transfert des valeurs contenues dans le message d'erreur filtré dans des phrases préformatées avec des variables à instancier. L'exemple d'un message de diagnostic d'une production erronée est affiché sur la figure 7.

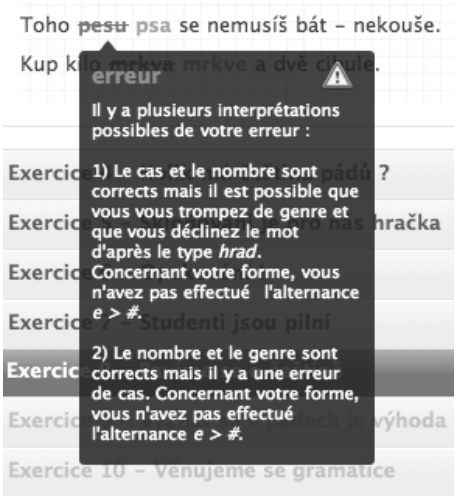

Fig. *7. Exemple d'un message de diagnostic*

5. Évaluation

Nous avons menée une étude pilote visant à tester le dispositif avec des données authentiques recueillies par deux enquêtes différentes. Il s'agit d'une illustration des possibilités de CETLEF dont l'objectif principal est de montrer des exemples d'enquêtes qui sont menées grâce à cet outil et qui peuvent être utilisées pour une recherche sur l'acquisition de la déclinaison du

tchèque par les francophones.

Dans l'enquête publique, qui a eu lieu sur www.cetlef.fr, 159 exercices ont été envoyés à la correction au cours d'une période de deux mois. Au sein de ces exercices, 1551 tâches ont été effectuées. Le nombre de productions erronées dans ces tâches est égal à 442 (28,5 % de toutes les productions). Pour 61 productions erronées (13,8 % de toutes les productions erronées), le diagnostic automatique n'a pas réussi à identifier la nature de l'erreur. Une représentation schématique de cette situation est proposée sur la figure 8.

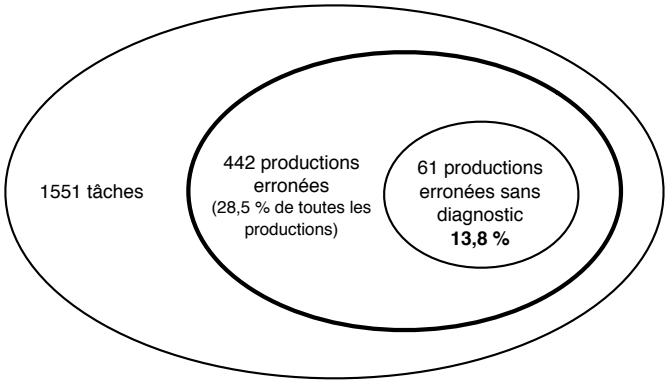

Fig. *8. Pourcentage de productions erronées et sans diagnostic*

Parmi toutes les productions erronées, il y en a 373 (84,4 %) qui ont été diagnostiquées comme une erreur avec une ou plusieurs interprétations morphologiques. Le nombre de productions erronées, diagnostiquées comme un des types d'erreurs établies d'après les restrictions sur les formes hypothétiques (ou une combinaison possible de deux types différents), est présentée dans la table 4.

5.0.5. Erreurs locales

Dans l'enquête publique, les erreurs les plus fréquentes parmi les erreurs locales sont les **erreurs de diacritique** (93 productions erronées, 75,6 %). L'explication de ce fait devrait être cherché dans les raisons techniques de saisie des caractères diacrités.

Les **erreurs d'alternance** sont le second type d'erreurs locales le plus fréquent (20 productions erronées, 16,3 %). Il est intéressant de constater que ce sont uniquement les formes requises avec les alternances vocaliques quantitatives qui étaient à l'origine de ces erreurs (*kráv* au lieu de *krav*, *jmén* au lieu de *jmen*, *díl* au lieu de *děl*, *houb* au lieu de *hub*, *můžem* au lieu de *nožem*, *penízmi* au lieu de *penězmi*, *smlouv* au lieu de *smluv*, etc.).

type de l'erreur	*productions*	%
erreur locale	123	33,0
erreur verticale	145	38,9
erreur horizontale interne	2	0,5
erreur horizontale externe	19	5,1
erreur locale ou verticale	39	18,5
erreur verticale ou horizontale interne	1	0,2
erreur verticale ou horizontale externe	40	10,7
erreur horizontale interne ou externe	4	1,1
sans diagnostic	61	

Tab. *4. Pourcentage des types d'erreurs*

5.0.6. Erreurs verticales

Parmi les erreurs verticales, les productions les plus fréquentes sont les **erreurs de cas** (71 productions, 49,0 %) qui reflètent probablement des dysfonctionnements au niveau syntaxique. Les erreurs les plus courantes sont celles où la production erronée a été laissée au nominatif singulier.

Le second groupe d'erreurs verticales les plus fréquentes sont les **erreurs de nombre** (19 productions avec un remplacement d'une forme de pluriel par une forme de singulier ; 8 productions avec un remplacement d'une forme de singulier par une forme de pluriel).

5.0.7. Erreurs horizontales externes

Contrairement aux erreurs horizontales internes (seulement deux occurrences), le diagnostic des erreurs horizontales externes se montre plutôt satisfaisant. Les plus nombreuses sont les **erreurs de type paradigmatique et de genre**, causées par la confusion du type paradigmatique en fonction des ambiguïtés qui peuvent exister dans l'attribution de ce type à un mot en fonction de sa forme. Par exemple, dans les tâches (5) et (6), il s'agit d'une confusion entre les types *předseda* et *žena*.

(5) *Bez* ****turist*** *je v Praze klid*
turistů gen.sg.m
sans touristes est à Prague calme

'Sans touristes, Prague est calme'

Diagnostic : *Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type* žena.

(6) *Hanička se líbí* ****Jirce****.*
Jirkovi gen.sg.f
Hanička *refl* plaît à Jirka

'Hanička plaît à Jirka'

Diagnostic : *Le cas et le nombre sont corrects mais peut être que vous vous trompez de genre et que vous déclinez le mot d'après le type* žena.

5.0.8. Erreurs locales ou verticales

La majorité des productions erronées pour lesquelles le diagnostic propose soit une interprétation locale, soit une interprétation verticale, sont des **erreurs de diacritique** en ce qui concerne l'interprétation locale. Le diagnostic peut être jugé comme adéquat si la seconde interprétation – l'interprétation verticale – est **une erreur de cas ou de nombre** sans contenir une erreur de diacritique.

5.0.9. Erreurs non morphologiques

Les productions qui ont été diagnostiquées comme n'ayant aucune interprétation morphologique mais qui remplissent cependant une des conditions posées dans les tests non morphologiques sont les suivantes : la production *pesmrkev* au lieu de *pes* a été diagnostiquée comme trop longue. Il s'agit évidemment d'une inattention, les productions de deux tâches distinctes ont été saisies dans un seul champ de formulaire sur la plateforme apprenant.

La production *rad* au lieu de *radost* a été diagnostiquée comme trop courte. Il s'agit ici probablement d'une confusion avec l'adjectif nominal *rád*, figurant dans la locution *Jsem rád, že ...* (*Je suis content que ...*), et qui remplace dans cette tâche la forme casuelle appropriée du substantif *radost*.

Deux productions ont été éliminées du traitement morphologique grâce au test sur la distance de Levehnstein entre la forme requise et la production erronée. Il s'agit des productions *houbovych* au lieu de *hub* et *vejcatach* au lieu de *vajec*.

5.0.10. Erreurs sans diagnostic

Les productions qui n'ont pas été diagnostiquées parmi les types d'erreurs présentés ci-dessus représentent 13,8 % de toutes les productions erronées. La plus grande partie de ces productions contient des fautes de frappe, manifestées le plus souvent par un ajout, un remplacement ou un effacement d'un graphème dans le radical de la production erronée. Cette modification rend le radical distinct par rapport au radical de la forme requise mais cette différence n'est pas assez prononcée pour que la production erronée puisse être diagnostiquée dans les traitements non morphologiques. Il s'agit par exemple de la production *spolužci* au lieu de *spolužáci*, *mínosti* au lieu de *místnosti*, *pkoje* au lieu de *pokoje*, *studenky* au lieu de *studentky*, *písnchí* au lieu de *písní*, etc.

6. Conclusion et perspectives

Nous estimons que l'apport principal de notre travail est l'intégration d'une riche représentation morphologique dans un outil d'enseignement de langue assisté par ordinateur.

Notre hypothèse que les erreurs de déclinaison sont calculables a été éprouvée dans le diagnostic automatique. L'application du diagnostic sur un échantillon de productions authentiques, collectées sur CETLEF, a permis de vérifier que cette hypothèse est vraie pour une grande partie de productions. La majorité des erreurs peut être interprétée automatiquement comme une combinaison du radical de la forme requise et d'une désinence.

Pour l'évaluation globale du diagnostic automatique, il est nécessaire de considérer la situation spécifique dans laquelle les erreurs analysées ont été produites. En déclinant une forme au sein d'un exercice grammatical, l'apprenant et la machine procèdent effectivement d'une manière assez semblable au niveau de l'analyse et de la génération. De ce point de vue, il serait intéressant d'étudier les occurrences des erreurs décrites dans ce travail dans les productions libres où l'apprenant n'est pas limité à un cadre défini aussi strictement que dans la tâche d'un exercice.

L'utilité du diagnostic pour la recherche des différents types d'erreurs dans la base de données est indéniable. Grâce à l'annotation morphologique des formes requises et grâce au message d'erreur caractérisant les productions erronées, des recherches basées sur cette annotation peuvent être effectuées facilement et servir des analyses variées, comme nous l'avons illustré avec un échantillon de données recueilli sur CETLEF. Avec le nombre croissant de productions dans la base de données au cours du temps, il sera possible d'entreprendre des études d'une envergure plus grande.

L'adéquation du diagnostic du point de vue didactique est une question qui reste ouverte pour le moment. Des modifications du diagnostic se révéleront nécessaires au fur et à mesure du service de CETLEF, avec le volume croissant de différentes productions erronées. Comme une des perspectives, il serait souhaitable de l'améliorer par l'intégration d'informations syntaxiques qui permettraient d'enrichir les critères pour le choix de l'interprétation la plus probable, pour identifier automatiquement des erreurs d'accord, des erreurs dans l'attribution d'une rection à un mot, etc.

L'outil CETLEF permet des analyses de volumes de données plus importantes que celles qui ont été exploitées dans notre travail. Ceci devrait permettre d'effectuer une analyse des erreurs dans l'acquisition de la déclinaison, qui serait menée non pas uniquement sur la base de critères morphologiques, comme c'est le cas dans notre travail, mais qui pourrait prendre en compte des critères plus complexes, comme les interférences entre les deux langues.

Remarque Cet article est le résumé de la thèse (Šmilauer, 2008), élaborée dans le cadre d'un doctorat en cotutelle entre le laboratoire LALIC-CERTAL (Langues, Logiques, Informatique, Cognition – Centre de Recherche en Grammaire et Traitement Automatique des Langues) de l'INALCO (Institut National des Langues et Civilisations Orientales) à Paris, et le laboratoire ÚTKL (Institute of Theoretical and Computational Linguistics) de la Faculté des Lettres de l'Université Charles de Prague.

Bibliographie

Bautier-Castaing, Elisabeth. 1977. Acquisition comparée de la syntaxe du français par des enfants fran-

cophones et non francophones. Étude expérimentale de quelques stratégies d'apprentissage. *Étude de linguistique appliquée*, 27 :19–41.

Bertin, Jean-Claude. 2001. *Des outils pour des langues. Multimédia et Apprentissage.* Ellipses Éditions, Paris.

Besse, Henri and Rémy Porquier. 1991. *Grammaires et didactiques des langues.* Hatier / Didier, Paris.

Cameron, Keitt, editor. 1999. *CALL : Media, Design and Applications*. Swets & Zeitlinger, Lisse.

Čermák, František and Michal Křen. 2004. *Frekvenční slovník češtiny.* Nakladatelství Lidové Noviny, Praha.

Gaonac'h, Daniel. 1991. *Théories d'apprentissage et acquisition d'une langue étrangère.* Hatier / Didier, Paris.

Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson. 2002. *Computer learner corpora, second language acquisition and foreign language teaching.* Benjamins, Amsterdam.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection. Computational Morphology of Czech.* Karolinum, Praha.

Hanson-Smith, Elizabeth. 2003. A brief history of CALL theory. *CATESOL Journal*, 15(1) :21–30.

Heift, Trude and Mathias Schulze. 2003. Error diagnosis and error correction in CAL : Introduction. *CALICO*, 20(3) :433–436.

Heift, Trude and Mathias Schulze. 2007. *Errors and Intelligence in Computer-Assisted Language Learning : Parsers and Pedagogues*. Routledge, UK.

Holland, V. Melissa and Jonathan D. Kaplan. 1995. NLP techniques in CALL : Status and instructional issues. *Instructional Science*, 23 :352–380.

Hrdlička, Milan. 2002. *Cizí jazyk čeština.* ISV, Praha.

Karttunen, F. 1986. A linguist looks at computer-assisted instruction. In Reinhold Freudenstein and James C. Vaughan, editors, *Confidence Through Competence in Modern Language Learning. CILT Reports & Papers 25*.

Kraif, Olivier, Georges Antoniadis, Sandra Echinard, Mathieu Loiseau, T. Lebarbé, and Claude Ponton. 2004. NLP tools for CALL : the simpler, the better. In *Proceedings of InSTIL / ICALL2004 – NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, 17-19 June.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707–710.

Levy, Michael. 1997. *Computer-Assisted Language Learning : Context and Conceptualization.* Clarendon Press, Oxford.

L'haire, Sébastien and Anne Vandeventer-Faltin. 2003. Diagnostic d'erreurs dans le projet FreeText. *ALSIC : Apprentissage des Langues et Systèmes d'Information et de Communication*, 6(2) :21–37.

Nekula, Marek. 2007. Systém a úzus. K výuce české deklinace se zřetelem k substantivům. In Jana Čemusová and Lída Holá, editors, *Sborník Asociace učitelů češtiny jako cizího jazyka 2006-2007*. Akropolis, Praha, pages 23–47.

Nerbonne, John. 2003. Natural language processing in computer-assisted language learning. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pages 670–698.

Nerbonne, John, Sake Jager, and Arthur van Essen, editors. 1998. *Language Teaching and Language Technology*. Swets & Zeitlinger, Lisse.

Osolsobě, Klára. 1996. *Algoritmický popis formální morfologie a strojový slovník češtiny*. Ph.D. thesis, Filozofická fakulta Masarykovy univerzity, Brno.

Poldauf, Ivan and Karel Špruňk. 1968. *Čeština jazyk cizí*. Státní pedagogické nakladatelství, Praha.

Porquier, Rémy. 1977. L'analyse des erreurs. Problèmes et perspectives. *Étude de linguistique appliquée*, 25 :23–43.

Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal*, 26 :81–114.

Šmilauer, Ivan. 2008. *Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison*. Ph.D. thesis, FF UK, INALCO, Prague, Paris.

Tono, Yukio. 2003. Learner corpora : design, development and applications. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference*, pages 800–809.

Tschichold, Cornelia. 2006. Intelligent CALL : The magnitude of the task. In P. Mertens, C. Fairon, A. Dister, and P. Watrin, editors, *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*, pages 806–814, Louvain-la-Neuve. Presses universitaires de Louvain.

Zock, Michael. 1996. Computational linguistics and its use in real world : The case of computer assisted-language learning. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 1002–1004.