Exploitation d'un outil ELIAO dans la recherche sur l'acquisition de L2

Ivan Šmilauer

Exploitation d'un outil ELIAO dans la recherche sur l'acquisition de L2

Ivan ŠMILAUER

INALCO, LaLIC-CERTAL
49bis, avenue de la Belle Gabrielle - 75012 Paris, FRANCE
Courriel : smilauer@cetlef.fr

ABSTRACT

ICALL (Intelligent Computer Assisted Language Learning) uses natural language processing for dealing with learner language. In this paper, we illustrate how to explore ICALL tools as a source of data for SLA studies on acquisition of inflectional morphology in written language. The platform CETLEF.fr is a dynamic Web application containing fill-in-the-blank exercises on Czech declension with an automatic error diagnosis and feedback. All data provided by learners and those generated automatically by the diagnosis algorithm are stored in a relational database which offers large possibilities of making queries regarding to the learner profile, the morphological properties of required forms in the declension tasks and the error annotation.

1. INTRODUCTION

Dans la recherche sur l'acquisition de L2, les ordinateurs sont utilisés pour la collecte des données, leurs stockage, structuration, visualisation et analyse (voir p.ex. Chapelle [Cha01]).

Nous présentons un outil informatique dont l'utilité première n'est pas de servir le chercheur mais l'objet de son investigation - l'apprenant. En effet, les *outils d'enseignement des langues "intelligemment" assisté par ordinateur* (ELIAO) sont conçus d'abord pour assister l'apprenant dans son apprentissage. Sur l'exemple de la plateforme CETLEF.fr[1], nous voulons montrer la possibilité d'utiliser ces dispositifs également comme moyen de constitution d'une riche base empirique pour l'analyse.

2. QU'EST-CE QUE L'ELIAO ?

Il s'agit d'un sous-ensemble des outils ELAO qui ne se satisfont pas seulement des nouvelles technologies pour reproduire les contenus et les activités traditionnelles, mais qui intègrent également les techniques du *traitement automatique des langues* (TAL) et plus largement, de l'*intelligence artificielle* (IA).

Ces techniques sont employées d'une part pour (1) *l'enrichissement des ressources* mises à la disposition des apprenants, d'autre part pour (2) *l'analyse de leurs productions* et un *feedback automatique personnalisé*.

[1] *Connaître, Comprendre et Corriger les Erreurs en Tchèque Langue Étrangère pour les Francophones* (CETLEF), disponible librement sur `http://www.cetlef.fr`. Il s'agit de la partie appliquée de Šmilauer [Šmi08].

2.1. Enrichissement des ressources

L'enrichissement des ressources pédagogiques peut être effectué de deux façons : (1) utilisation des outils génériques du TAL pour des fins pédagogiques (lexiques électroniques, outils de génération des formes fléchies, analyse et synthèse vocale, etc.) ; (2) emploi des outils de bas niveau (lemmatiseurs, analyseurs morphologiques ou syntaxiques) pour l'ajout de l'*annotation linguistique* (catégories lexicales et morphologiques, types de flexion etc.) aux divers contenus en L2.

Concernant la seconde approche, l'annotation des textes pédagogiques peut être employée pour la génération automatique des exercices ciblés sur un certain phénomène (p.ex. accord des adjectifs, conjugaison, déclinaison, prépositions, etc.), pour la visualisation explicite de cette information ou pour la constitution des lexiques personnalisés, structurés en fonction de ce paramètre.

D'autre part, l'annotation des productions des apprenants est la condition nécessaire pour l'implémentation d'un diagnostic automatique des erreurs qui génère un retour personnalisé.

Les deux types d'annotation peuvent être utilisés pour la caractérisation des données linguistiques collectées pour les objectifs d'une analyse.

2.2. Diagnostic des erreurs et feedback

Les projets d'un système automatique qui assisterait l'apprenant dans son apprentissage de L2 en réagissant d'une manière appropriée à ces productions proviennent des travaux en intelligence artificielle des années 80.

Dans cette perspective, l'enjeu pour ELIAO n'est pas uniquement d'identifier et de corriger les erreurs mais également de les diagnostiquer afin de générer un feedback. Cette sorte de retour automatique devrait amener l'apprenant à comprendre son erreur et essayer de l'éviter à l'avenir (voir Heift & Schulze [Hei07] ou Meurers [Meu09]).

Comme pour les autres tâches liées au traitement automatique du langage naturel, les objectifs, ambitieux à l'origine, se sont vite réduits à la résolution partielle des problèmes fondamentaux. Il semble qu'il est plus convenable de s'appuyer sur des techniques relativement bien maîtrisées (calculs sur les chaînes de caractères, analyse morphologique), car la fiabilité actuelle des méthodes plus sophistiquées (analyse syntaxique ou sémantique) ne permet pas d'obtenir des résultats satisfaisants en éliminant

le bruit et les erreurs qui pourraient avoir un effet néfaste sur l'apprenant.

Nous allons montrer que le diagnostic basique des erreurs morphologiques peut être utilisé non seulement pour un feedback destiné à l'apprenant mais également pour la spécification des formes erronées dans une base de données (BD).

3. PRÉSENTATION DE CETLEF.FR

Notre outil est une application web dynamique qui propose des exercices de déclinaison tchèque. La tâche unique de l'apprenant sur cette plateforme est de décliner une forme donnée (substantif, adjectif ou pronom) dans son contexte syntaxique au sein d'une phrase à trous. La structure d'une telle tâche ainsi que la terminologie employée sont présentées sur la figure 1.

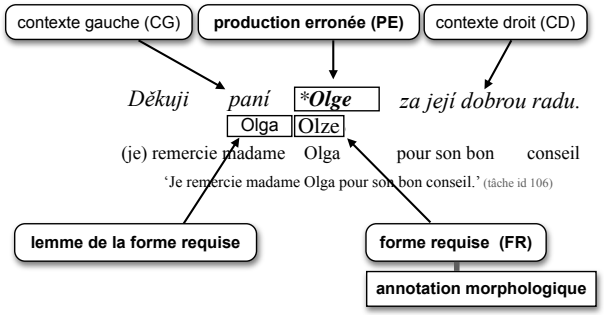

Figure 1: Une tâche de déclinaison avec la terminologie employée.

3.1. Étude de la flexion nominale

Le choix de la déclinaison tchèque comme matière traitée sur une plateforme ELIAO et comme objet de recherche en acquisition de L2 peut être justifié par la richesse de sa flexion nominale (4 genres, 7 cas et 14 types de base de déclinaison des substantifs). Cette complexité est un élément important à surmonter pour les apprenants, notamment pour ceux dont la langue maternelle ne possède pas de morphèmes spécifiques pour exprimer la catégorie de cas.

Plus largement, l'acquisition et le traitement de la morphologie flexionnelle est un sujet récurrent dans les travaux menés dans la perspective psycho-computationnelle (voir p.ex. Pirelli [Pir07]) qui tentent de modéliser ces processus à l'aide des formalismes explicites. Les questions posées dans ces travaux concernent le rôle et la place des connaissances déclaratives (lexique mental) et procédurales (règles morphologiques et syntaxiques) dans la production linguistique et elles s'approchent des modélisations dans le TAL et IA.

3.2. Architecture du dispositif

La plateforme est une application web dynamique (programmée en PhP et HTML) avec une BD relationnelle MySQL. Elle contient deux parties principales :

La **plateforme auteur** permet de concevoir les exercices à l'aide d'une interface graphique ergonomique. Pour chaque tâche, l'auteur saisit (1) le contexte gauche et droit de la forme requise, (2) la *forme requise* (FR) et (3) l'annotation de la FR[2]. Cette annotation, dont les attributs morphologiques sont présentés sur la figure 2, est proposée à l'auteur automatiquement à l'aide d'un analyseur morphologique rudimentaire.

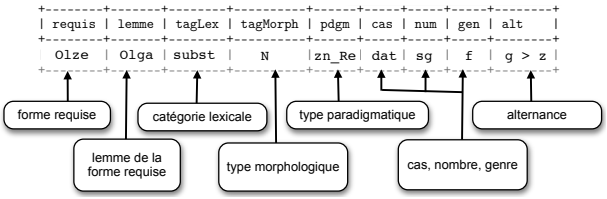

Figure 2: Spécification des attributs dans l'annotation d'une forme requise.

La **plateforme apprenant** est ouverte à chacun qui s'inscrit en tant qu'utilisateur, en fournissant des données permettant d'établir son profil sociolinguistique (sexe, âge, durée d'apprentissage du tchèque, suivi pédagogique, autres langues maîtrisées, etc.). Les exercices sont proposés dans un ordre fixe, établi par l'auteur. Au sein de chaque exercice, exécuté avec une limite de temps, l'apprenant accomplit les tâches de déclinaison en saisissant une *production correcte* (égale à FR) ou une *production erronée* (PE). Lors de la correction des exercices, l'apprenant peut consulter l'annotation de la FR et s'il a commis une erreur, également le feedback sur sa production.

4. CARACTÉRISTIQUES DES PRODUCTIONS

Afin de situer les données collectées sur CETLEF.fr, nous présenterons les différents types de productions utilisées dans les recherches sur L2 d'après leur source et leur matérialité.

4.1. Source des productions

En fonction des paramètres situationnels, on distingue traditionnellement (voir Ellis [Ell94]) : (1) les productions dites *naturelles*, produites dans une situation communicative non-artificielle (p.ex. des productions dans un contexte authentique) ; (2) les productions *sollicitées* dans le cadre d'une expérimentation (p.ex. des productions dans des tâches ciblées sur un phénomène grammatical).

Les productions sollicitées doivent nécessairement contenir les éléments recherchés par l'expérimentateur, ce qui est leur avantage. En revanche, un manque d'authenticité peut leur être reproché. Ellis & Brakhuizen [Ell05] soulignent que les lacunes dans l'*analyse des erreurs* classique, travaillant avec des textes sujets à des stratégies d'évitement, peuvent être complétées par des études expérimentales.

Dans le cas des productions sollicitées, l'opposition entre celles qui sont issues des tâches *fermées* (nombre limité de productions possibles, p.ex. les phrases à trous ou une traduction simple) ou *ouvertes* (nombre de productions virtuellement illimité, p.ex. dans une rédaction) doit être également prise en compte. Cette opposition est cruciale pour

[2]Nous envisageons de simplifier cette procédure par la lemmatisation de la FR dans une phrase authentique, choisie dans un corpus annoté.

le choix des moyens mobilisés pour l'analyse automatique des erreurs car elle détermine les restrictions sur les interprétations possibles d'une production erronée.

4.2. Données numérisées et leur accessibilité

Dans les recherches sur L2, les productions numérisées sont exploitées notamment au sein des *corpus d'apprenants*, voir notamment Granger [Gra09]. Une annotation morphologique qui permettrait d'inclure des critères métalinguistiques dans les requêtes sur ces corpus, est plutôt rare, notamment à cause des problèmes liés à l'application des outils génériques à des textes comportant des erreurs et des irrégularités.

L'étape de la numérisation "manuelle" des données, inévitable pour la compilation d'un corpus de productions libres, peut être supprimée par la collecte des productions qui sont saisies sur l'ordinateur par l'apprenant lui-même[3].

Tandis que les productions libres sont stockées dans un corpus et exploitées par un concordancier, les productions issues des tâches fermées sur une plateforme ELIAO peuvent être facilement insérées dans une base de données. Le langage d'interrogation des BD et les informations sur les apprenants, les tâches et les productions provenant de la plateforme permettent la formulation des requêtes complexes et précises, ce qui peut être d'une grande utilité pour l'analyse.

5. DIAGNOSTIC DES ERREURS

Les données recueillies sur CETLEF.fr sont donc des productions sollicitées dans des tâches fermées qui sont annotées morphologiquement. Ce cadre a permis l'implémentation d'un diagnostic automatique basé sur la génération morphologique, appuyé par un modèle formel de déclinaison tchèque élaboré à cet effet.

5.1. Hypothèse sur les erreurs

Par hypothèse, la forme d'une *production erronée* (PE) n'est pas considérée comme aléatoire mais comme le résultat d'une activité succombant à des règles d'ordre cognitif. Nous supposons qu'elle peut être calculée à partir du lemme de la forme requise (FR) à l'aide de (1) l'ajout des désinences inappropriées (au sein du paradigme correspondant mais également au sein des autres), (2) le non respect des alternances morphématiques et (3) le marquage incorrect de la diacritique[4]. Une forme qui peut être calculée de cette manière est appellée une *erreur morphologique*.

5.2. Algorithme de diagnostic

Après avoir éliminé les formes improbables (longueur inadéquate, distance de Levehnstein trop importante, etc.), l'algorithme de diagnostic génère pour chaque PE un ensemble de *formes hypothétiques* (FH) à partir du lemme de la FR. P.ex. pour la FR *Olze*, les PH sont *Olga, Olgy, Olge, Olgu, ..., Olgovi, ... Olgům, ...*

[3]Les différents aspects de la collecte des données par ce moyen sont présentés dans Hulstijn [Hul00].

[4]Une méthode similaire a été employée par Kraif & Ponton [Kra07] pour l'analyse des erreurs d'un corpus de productions libres d'apprenants du français.

Si une correspondance entre une des FH et la PE est trouvée, la différence entre les valeurs des attributs morphologiques de la FH et de la FR est retenue comme une interprétation possible de la PE (voir la figure 3).

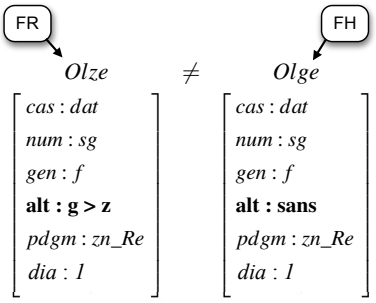

Figure 3: La comparaison de la forme requise (FR) avec une des formes hypothétiques (FH) donnant l'interprétation *erreur d'alternance*.

5.3. Types des erreurs morphologiques

En fonction des attributs morphologiques atteints par l'erreur, une PE peut être donc interprétée comme une *erreur de cas, de genre, de nombre, de type paradigmatique, d'alternance* et *de diacritique* ou la combinaison de ces valeurs.

Plusieurs interprétations différentes sont souvent retenues pour une PE car la plupart des désinences casuelles tchèques sont homonymes. Dans cette situation, un algorithme de filtrage ne retient que les deux interprétations qui diffèrent de la FR dans le plus petit nombre d'attributs.

Le diagnostic filtré est par la suite enregistré dans la BD dans un message d'erreur. Pour la PE sur la figure 1, ce message a la forme suivante :

```
§ LOC|zn_Re|e|dat|sg|f|g>z|ERR
```

Pour le message ci-dessus, le message de feedback destiné à l'apprenant est le suivant :

Aucune erreur de cas, de genre ni de nombre mais vous n'avez pas effectué l'alternance g > z.

5.4. Limites du diagnostic

Avec les informations disponibles dans l'annotation de la FR, le diagnostic de la PE reste limité au niveau morphologique. Pour les PE avec une désinence homonyme, il est impossible de décider entre certaines interprétations sans prendre en compte des informations d'ordre syntaxique ou sémantique.

Pour illustration, une PE *růží* (lemme *růže*, f. "rose") au lieu de *růže* (génitif sg.), peut être interprétée indépendamment du contexte comme le génitif pl. (erreur de nombre) ou l'instrumental sg. (erreur de cas). Le nombre égal d'attribut atteints par l'erreur ne nous permet pas de privilégier une de ces interprétations : (1) l'apprenant est-il conscient de la valeur du cas de la forme requise (exigée par sa fonction syntaxique) et il confond seulement le paradigme pluriel avec le singulier ; ou (2) choisi-t-il une désinence différente au sein du paradigme du singulier, parce qu'il fait une hypothèse erronée sur la position

syntaxique de la FR ?

Dans la situation où la FR serait un argument d'un verbe qui peut avoir les deux rections possibles (génitif et instrumental), la deuxième hypothèse serait plus probable. La situation serait encore plus claire si l'équivalent du verbe tchèque dans la langue maternelle de l'apprenant exigerait un instrumental. Si nous disposions d'une telle information, le diagnostic pourrait être perfectionné, il s'avère néanmoins que la désambiguïsation des PE est une opération délicate même pour une analyse "manuelle".

6. ÉVALUATION

CETLEF.fr a été mis en service en juin 2008. Dans le cadre d'une enquête préliminaire, destinée prioritairement aux tests et aux évaluations du dispositif, 29 exercices (291 tâches) ont été proposés aux apprenants. Ces exercices sont ordonnés d'après la progression des cas dans les matériaux pédagogiques courants destinés aux débutants. Les formes requises ont été choisies de manière à ce qu'elles représentent la majorité des paradigmes et des alternances morphématiques qui y sont présentés.

6.1. Données recueillies

Au 20 novembre 2009, 164 apprenants étaient inscrits sur la plateforme, fournissant au total 6667 productions, dont 2535 (38 %) sont des PE. Etant donné que les exercices sont accessibles l'un après l'autre, toutes les tâches étaient accomplies uniquement par les 13 apprenants les plus persévérants.

Parmi toutes les PE, 2008 productions (79 %) ont été diagnostiquées en tant qu'erreur morphologique ce qui est un résultat encourageant pour l'algorithme du diagnostic ainsi que pour l'hypothèse sur la calculabilité des erreurs. 5 % des PE ont été interprétées par des méthodes non-morphologiques et les 16 % restant représentent le défi pour l'amélioration du diagnostic, car à l'heure actuelle, aucune interprétation n'est trouvée pour ces formes.

Etant donné le caractère provisoire de l'enquête préliminaire, il ne serait pas utile de publier ici des chiffres exacts concernant les différents types d'erreurs. Un apperçu général montre que les erreurs les plus fréquentes sont les erreurs de cas ($\pm$ 40 %), suivies des erreurs de diacritique ($\pm$ 20 %), de genre ($\pm$ 15 %), de nombre ($\pm$ 15 %) et d'alternance ($\pm$ 5 %). Les erreurs restantes sont les combinaisons diverses des types cités ci-dessus.

6.2. Potentiel pour l'analyse de L2

Les productions peuvent être recherchées et triés en fonction des données contenues dans (1) le profil de l'apprenant, (2) l'annotation morphologique de la FR et (3) le message d'erreur caractérisant la PE.

Grâce aux outils annexes exploitant la BD et permettant une visualisation synoptique des résultats dans un navigateur web de base, il est possible d'étudier toutes les productions dans une tâche (voir figure 4), toutes les productions d'un apprenant, toutes les occurrences d'un type d'erreur, toutes les occasions de produire une certaine forme ou d'effectuer une certaine opération dans des tâches différentes, etc.

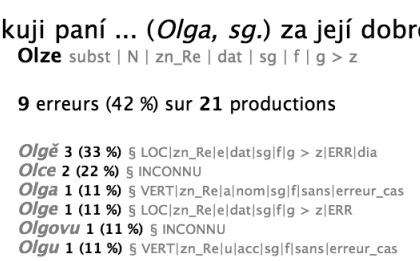

Figure 4: Exemple de visualisation des PE pour une tâche.

6.3. Conclusion et perspectives

Bien que les données traitées par CETLEF.fr soient des productions écrites issues des exercices sur la déclinaison tchèque, le concept général illustré par notre plateforme peut être employé pour d'autres langues et sur d'autres problèmes linguistiques à condition que la limite de la morphologie ne soit pas dépassée.

Nous espérons avoir montré qu'une procédure de diagnostic relativement simple et une description formelle des erreurs peut faciliter le travail du chercheur dans les analyses de L2 plus approfondies.

RÉFÉRENCES

[Cha01] Chapelle, C.A. (2001), *Computer Applications in Second Language Acquisition*, Cambridge, UK : Cambridge University Press.

[Ell94] Ellis, R. (1994), *The Study of Second Language Acquisition*, Oxford University Press.

[Ell05] Ellis, R. and Barkhuizen, G. (2005), *Analysing learner language*, Oxford University Press.

[Gra09] Granger, S. (2009), "The contribution of learner corpora to SLA and FLT : A critical evaluation", in *Corpora and Language Teaching*, K. Aijmer, ed., Benjamins, pp. 13–23.

[Hei07] Heift, T. and Schulze, M. (2007), *Errors and Intelligence in Computer-Assisted Language Learning : Parsers and Pedagogues*, UK : Routledge.

[Hul00] Hulstijn, J.H. (2000), "The use of computer technology in experimental studies of second language acquisition", *Language Learning and Technology*, vol. 3, pp. 32–43.

[Kra07] 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).

[Meu09] Meurers, D. (2009), "On the Automatic Analysis of Learner Language. Introduction to the Special Issue", *CALICO Journal*, vol. 26, pp. 469–473.

[Pir07] Pirelli, V. (2007), "Psycho-computational issues in Morphology Learning and Processing : An ouverture", *Lingue e Linguagguo*, vol. VI, pp. 131–138.

[Šmi08] Šmilauer, I. (2008), *Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison*, Ph.D. thesis, INALCO, Paris / Faculté des lettres de l'Université Charles, Prague.