CETLEF.fr - diagnostic automatique des erreurs de déclinaison tchèque dans un outil ELAO

Ivan Šmilauer

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

Résumé CETLEF.fr – une application Web dynamique – propose des exercices de déclinaison tchèque avec un diagnostic automatique des erreurs. Le diagnostic a nécessité l'élaboration d'un modèle formel spécifique de la déclinaison contenant un classement des types paradigmatiques et des règles pour la réalisation des alternances morphématiques. Ce modèle est employé pour l'annotation des formes requises, nécessaire pour le diagnostic, mais également pour une présentation didactique sur la plateforme apprenant. Le diagnostic est effectué par comparaison d'une production erronée avec des formes hypothétiques générées à partir du radical de la forme requise et des différentes désinences casuelles. S'il existe une correspondance, l'erreur est interprétée d'après les différences dans les traits morphologiques de la forme requise et de la forme hypothétique. La majorité des erreurs commises peut être interprétée à l'aide de cette technique.

Abstract CETLEF.fr – a dynamic Web application – contains fill-in-the-blank exercises on Czech declension with an automatic error diagnosis. The diagnosis rendered necessary the definition of a specific formal model of nominal inflection containing a classification of the paradigms and the rules for the realization of morphemic alternations. This model has been employed for the morphological annotation of required forms, necessary for the error diagnosis as well as for a didactic presentation on the learning platform. Diagnosis is carried out by the comparison of an erroneous production with hypothetical forms generated from the radical of the required form and various haphazard endings. If a correspondence is found, the error is interpreted according to the differences in the morphological features of the required form and the hypothetical form. The majority of errors can be interpreted with the aid of this technique.

Mots-clés : morphologie flexionnelle, déclinaison tchèque, acquisition d'une langue étrangère, diagnostic des erreurs et feedback, ELAO

Keywords: inflectional morphology, Czech declension, second language acquisition, error diagnosis and feedback, CALL

1 Erreur de déclinaison sur CETLEF.fr

CETLEF.fr[1] est un outil d'enseignement de langue assisté par ordinateur (ELAO) proposant des exercices de déclinaison tchèque : la tâche de l'apprenant est de créer la forme fléchie d'un lemme en fonction de son contexte syntaxique au sein d'une proposition donnée. Une erreur commise dans une telle tâche est appelée *erreur de déclinaison*. Notre objectif a été de concevoir un diagnostic des productions erronées issues de ces exercices.

Une erreur de déclinaison est considérée non pas comme un phénomène aléatoire mais comme le résultat d'une activité succombant à des règles d'ordre linguistique et cognitif. Dans la perspective de la linguistique contrastive, les erreurs de déclinaison peuvent être expliquées par la comparaison du système nominal tchèque et français : la déclinaison représente pour tout apprenant francophone une sorte d'idiosyncrasie et elle est inévitablement une source d'erreurs.

L'hypothèse sous-jacente au diagnostic des erreurs est que les erreurs de déclinaison sont calculables : une forme erronée peut être générée à partir du lemme de la forme requise au sein d'une tâche à l'aide des opérations d'ordre linguistique. Ainsi, pour le diagnostic d'une forme erronée, nous proposons de prendre en compte une défaillance dans une ou plusieurs des opérations suivantes : (1) choix des valeurs de la catégorie du cas, du nombre et du genre ; (2) classement du lexème dans le paradigme approprié et le choix de la désinence correspondante aux valeurs des catégories grammaticales ; (3) réalisation des alternances vocaliques et consonantiques.

2 Diagnostic des erreurs

Les différentes modélisations de la déclinaison tchèque implémentées dans des applications de TAL existantes, cf. par exemple (Hajič, 2004), ne correspondaient pas à notre objectif à cause d'une orientation fonctionnelle sur un traitement informatique qui ne prenait pas en compte le processus « cognitif » de génération des formes casuelles envisagé par un humain. Nous avons donc élaboré un modèle formel contenant un classement des paradigmes, définis uniquement par les ensembles de morphèmes, et les règles pour la réalisation des alternances en fonction des propriétés morphématiques et phono-graphémiques du radical et de la désinence. Ce modèle sert comme la base d'une annotation spécifiant la catégorie lexicale, les catégories morphologiques, les types paradigmatiques de déclinaison et les alternances à réaliser pour chaque forme requise. L'annotation, attribuée à l'aide d'un étiqueteur rudimentaire, est également utilisée pour une présentation didactique sur la plateforme apprenant.

Au cours du traitement automatique, l'annotation de chaque forme requise est représentée par une structure de traits. Pour chaque forme requise, un ensemble de formes hypothétiques, qui pourraient être produites par un apprenant, est généré à partir de son lemme. Il s'agit : (1) des formes casuelles exprimant d'autres valeurs des catégories grammaticales que celles

---

[1] *Connaître, Comprendre, Corriger les Erreurs en Tchèque Langue Étrangère pour les Francophones*, disponible sur `http://www.cetlef.fr`. Pour plus de détails voir (Šmilauer, 2008). Les moyens techniques employés sont PhP, MySQL, Javascript, HTML et XML.

demandées dans la forme requise, (2) des formes qui sont le résultat de la combinaison du radical avec une désinence qui n'appartient pas à son paradigme et (3) les formes sans la réalisation des alternances vocaliques ou consonantiques obligatoires.

À chacune de ces formes hypothétiques peut être assignée au moins une structure de traits (son *interprétation*). Une production erronée peut avoir une *interprétation morphologique* si elle correspond au moins à une des formes hypothétiques générées pour le lemme de la forme requise donnée. Les attributs communs portant des valeurs différentes dans les structures des deux formes servent pour la définition du type d'erreur.

L'ensemble des différentes interprétations morphologiques d'une production erronée est déterminé par l'homonymie des formes casuelles existantes, mais également par toutes les combinaisons du radical et des désinences, appartenant même aux autres paradigmes, ainsi que par la réalisation ou non des alternances éventuelles. Le filtrage des différentes interprétations (seulement deux interprétations sont retenues pour le feedback destiné à l'apprenant) est effectué à l'aide des règles qui déterminent la priorité entre les différentes interprétations : par exemple, interpréter une production en tant qu'une erreur de nombre est considéré comme plus pertinent qu'une erreur de cas.

3 Évaluation et application de CETLEF.fr

Une étude visant à tester l'efficacité du diagnostic avec des productions authentiques recueillies depuis juin 2008 a approuvée notre hypothèse de départ : environ 85 % des erreurs peuvent être diagnostiquées par un calcul morphologique comme une combinaison illicite du radical de la forme requise et d'une désinence.

CETLEF.fr illustre les possibilités d'un riche modèle morphologique et des techniques de TAL employées dans un outil d'enseignement de langue assisté par ordinateur – une approche préconisée notamment par (Heift & Schulze, 2007). Dans une perspective pédagogique, un feedback personnalisé basé sur le diagnostic permet d'attirer l'attention des apprenants sur les éléments problématiques.

Du point de vue d'une recherche sur l'acquisition de langue étrangère, CETLEF.fr sert comme un outil pour la compilation d'un corpus d'erreurs. Par rapport aux productions libres, l'analyse des erreurs recueillies au sein des exercices grammaticaux permet un meilleur contrôle sur le volume de données pertinentes, car les productions doivent nécessairement contenir les phénomènes qui ont été établis comme l'objet de l'investigation. L'annotation morphologique et la typologie formelle des erreurs facilitent cette analyse.

Références

HAJIČ J. (2004). *Disambiguation of Rich Inflection. Computational Morphology of Czech.* Prague : Karolinum.

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

ŠMILAUER I. (2008). Acquisition du tchèque par les francophones : analyse automatique des erreurs de déclinaison. *The Prague Bulletin of Mathematical Linguistics* 90, 33-56.