Genre des substantifs en tchèque : l'ambiguïté de ses marqueurs formels et le diagnostic automatique des erreurs du point de vue de son acquisition par les apprenants francophones

Ivan Šmilauer

Genre des substantifs en tchèque

l'ambiguïté de ses marqueurs formels du point de vue de son acquisition par les apprenants francophones et le diagnostic automatique des erreurs

Ivan ŠMILAUER
LaLIC–CERTAL, INALCO (Paris)
smilauer@cetlef.fr

session poster COLDOC'09

Ambiguïté morphologique du point de vue d'un apprenant de LE

Les effets de **l'ambiguïté**, définie dans un modèle linguistique stratificationnel par la **multitude de fonctions prises par un élément de niveau n au niveau n+1** (cf. Sgall et al. 1986) peuvent être également étudiés dans les productions langagières des apprenants d'une langue étrangère. Nous l'illustrons par l'étude des phénomènes d'ambiguïté entre le niveau phonologico-graphémique et le niveau morphologique du tchèque qui résultent en différentes **erreurs de genre dans les productions des apprenants francophones**.

Dans la plupart des cas, la relation entre un substantif et son genre grammatical ne présente pour un locuteur natif aucun caractère d'ambiguïté car cette information fait partie de son stock « naturel » de connaissances lexicales. La situation est différente pour un apprenant étranger confronté à un substantif au genre inconnu. Celui-ci se trouve devant deux solutions : rechercher et intégrer l'information sur genre « en dur » dans son vocabulaire ou essayer de trouver et d'appliquer des règles basées sur la forme qui permettrait d'attribuer le genre sans recours à un lexique. En se servant de cette seconde technique, l'apprenant se trouve nécessairement heurté aux limites des règles établies dont la surgénéralisation produit des erreurs.

C'est ainsi que nous considérons une erreur de genre comme le résultat de l'attribution d'une fonction inadéquate à une forme ambiguë par rapport aux règles dont se sert l'apprenant. De ce point de vue, **nous estimons que les substantifs sont plus ou moins ambigus par rapport au genre**.

Le tchèque possède quatre valeurs pour la catégorie grammaticale de genre : masculin animé et inanimé, féminin, neutre. A l'aide des critères formels et sans recours à une connaissance d'ordre lexicale, **il est possible de définir des règles pour l'attribution du genre en fonction des marqueurs situés à la périphérie droite des lexèmes** (la désinence casuelle vocalique du lemme ou la terminaison consonantique du radical). Ces marqueurs permettent avec un degré variable de certitude d'identifier le genre des substantifs : le nominatif singulier terminé par "–o" désigne sans ambiguïté un substantif neutre, le nominatif singulier terminé par "–a" désigne le plus souvent un féminin, mais aussi, marginalement, un masculin etc.

Dans une production d'apprenant, une erreur de genre peut se manifester (a) par une **erreur d'accord** des éléments dépendants de ce substantif, (b) par l'**attribution d'un type paradigmatique erroné** qui se manifeste lors de la déclinaison du substantif en question. Par exemple, le substantif masculin "stůl" (table) peut être interprété par l'apprenant en tant que féminin car un substantif terminé par la consonne "–l" peut être masculin, mais également féminin, voir par exemple "sůl" (sel). Ainsi, cette analyse erronée peut mener à des productions du type (a), par exemple "velká stůl", qui pourrait être traduit littéralement comme « grand table » ; ou à des erreurs de type (b), par exemple le génitif singulier "stoli" d'après le type de déclinaison consonantique féminin correspondant, au lieu de "stolu" d'après le type correct masculin. C'est sur ce deuxième type d'erreur que nous nous concentrons ici.

L'étude de l'ambiguïté des marqueurs formels du genre est intéressante également du point de vue pédagogique car cette propriété est en relation directe avec le degré de **diagrammaticité** du substantif (caractère indexical de sa construction formelle par rapport à sa signification) qui peut faciliter le traitement cognitif du lexème par l'apprenant.

Nous apportons ici des éléments pour étudier, dans un cadre expérimental spécifique (exercices de déclinaison sur une plate-forme ELAO), les erreurs de genre produites par des apprenant francophones et le rôle de l'ambiguïté des marqueurs formels des substantifs. Nous supposons que les erreurs de genre les plus fréquentes devraient être celles qui occurrent dans les substantifs formellement ambigus par rapport à la catégorie du genre.

En nous basant sur une **étude quantitative**, nous présentons d'abord l'ambiguïté des marqueurs de genre en tchèque. Ensuite, nous introduisons l'application CETLEF.fr (un **outil d'enseignement du tchèque assisté par ordinateur avec un diagnostic automatique des erreurs**) qui sert pour la collecte et l'analyse des productions langagières des apprenants.

Finalement, nous présentons les **données recueillies** à l'aide de CETLEF.fr et nous proposons un **algorithme modélisant l'attribution du genre** à un substantif par les choix d'ordres formels, sémantiques et lexicaux effectués par l'apprenant pendant l'analyse de la forme au genre inconnu. Cet algorithme pourrait servir à des fins pédagogiques mais également pour l'affinement du diagnostic automatique existant au sein de CETLEF.fr.

Étude quantitative des marqueurs formels du genre

La définition d'un marqueur formel est en principe arbitraire – le marqueur ultime d'un lexème serait sa forme entière, ce qui revient à même que de construire un lexique.

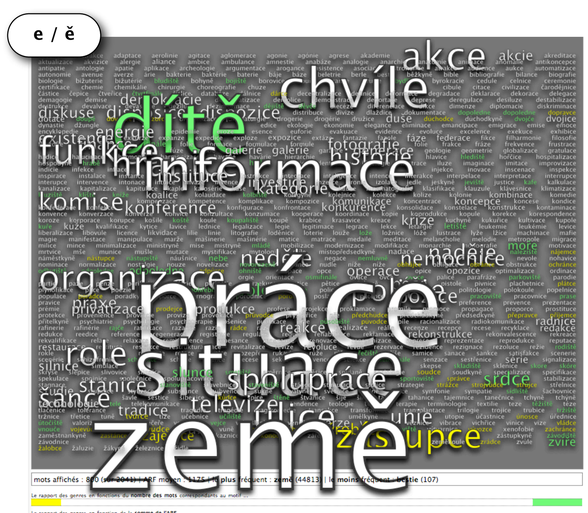
Etant donné le souci pour l'économie des moyens employés dans les règles, nous avons défini 7 marqueurs différents, couvrant uniquement le dernier phonème/graphème du mot : 4 marqueurs vocaliques (**o, í, a, e/ě**) et 3 classes de marqueurs consonantiques (**consonne dure, consonne mixte, consonne molle**).

Pour chaque marqueur, nous avons recherché les substantifs correspondants dans un lexique annoté contenant les 21 986 substantifs tchèques les plus fréquents dans le **Corpus National Tchèque** (cf. www.korpus.cz, Čermák 1997, Čermák & Křen 2004, 2005). Pour une meilleure visualisation des données qui permet une appréciation générale de l'ambiguïté des marqueurs, nous avons attribué à chaque mot une **couleur** en fonction de son genre (masculins animés : jaunes ; masculins inanimés : oranges ; féminins : blanc ; neutres : verts) et une **taille** en fonction de sa **fréquence moyenne réduite** (ARF – Average Reduced Frequency, cf. Hlaváčová and Savický 2002). Seulement les premiers 800 substantifs les plus fréquents sont affichés. Le rapport entre les différents genres pour un marqueur donné offre une possibilité de mesurer le taux de son ambiguïté.

Il serait possible d'établir des marqueurs à l'aide des segments plus importants, notamment ceux qui peuvent être identifiés comme des **suffixes de dérivation**. Néanmoins, cet affinement, qui va au détriment de la maniabilité des règles par l'apprenant, n'apporte pas forcément une baisse significative de leur ambiguïté.

A côté des critères formels, un moyen possible de lever l'ambiguïté d'un substantif est son analyse par rapport au **genre sémantique** (naturel) de l'objet désigné. Dans le cas des personnes et des animaux mâles et femelles, le genre grammatical et sémantique sont identiques, ainsi que pour les enfants humains et animaux (neutre). Dans le cas des objets, cette relation est complètement arbitraire. Une fois identifiée, la différence entre les masculins animés et inanimés est gérée par des critères du même ordre.

e / ě



mots affichés : 800 (sur 2041) | ARF moyen : 1175 | le plus fréquent : země (44813) | le moins fréquent : bytost (107)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 43 5.4 % de mots 5.5 % de ARF total ARF max : zástupce (10760) ARF moyen : 1191	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0	Nombre : 685 85.6 % de mots 84.8 % de ARF total ARF max : země (44813) ARF moyen : 1163	Nombre : 72 9 % de mots 9.8 % de ARF total ARF max : dítě (23934) ARF moyen : 1276

o

mots affichés : 752 (sur 752) | ARF moyen : 692 | le plus fréquent : místo (53526) | le moins fréquent : desolato (10)

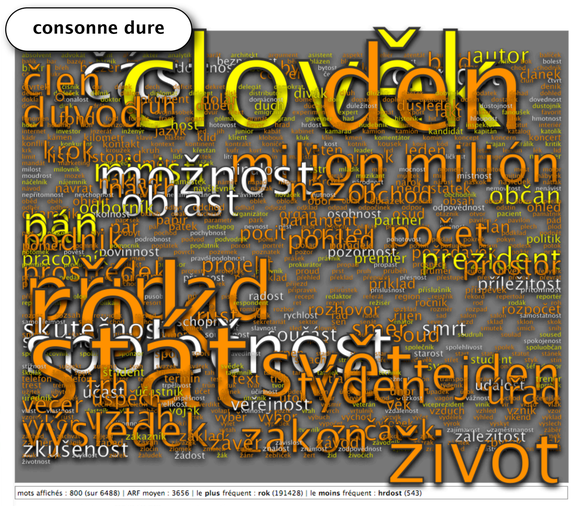
Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 4 0.5 % de mots 0 % de ARF total ARF max : maestro (33) ARF moyen : 19	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0	Nombre : 748 99.5 % de mots 100 % de ARF total ARF max : místo (53526) ARF moyen : 696

consonne dure



mots affichés : 800 (sur 6488) | ARF moyen : 3656 | le plus fréquent : rok (191428) | le moins fréquent : hrdost (543)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 149 18.6 % de mots 15.7 % de ARF total ARF max : člověk (83780) ARF moyen : 3079	Nombre : 549 68.6 % de mots 71 % de ARF total ARF max : rok (191428) ARF moyen : 3782	Nombre : 102 12.8 % de mots 13.3 % de ARF total ARF max : společnost (33502) ARF moyen : 3817	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0

í

mots affichés : 800 (sur 3981) | ARF moyen : 1052 | le plus fréquent : jednání (13861) | le moins fréquent : opovržení (178)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0	Nombre : 5 0.6 % de mots 1.8 % de ARF total ARF max : mluvčí (7185) ARF moyen : 3055	Nombre : 795 99.4 % de mots 98.2 % de ARF total ARF max : jednání (13861) ARF moyen : 1040

consonne mixte

mots affichés : 800 (sur 2095) | ARF moyen : 1431 | le plus fréquent : čas (31428) | le moins fréquent : baseball (114)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 136 17 % de mots 13.2 % de ARF total ARF max : ředitel (16468) ARF moyen : 1110	Nombre : 565 70.6 % de mots 77.4 % de ARF total ARF max : čas (31428) ARF moyen : 1567	Nombre : 31 3.9 % de mots 3.6 % de ARF total ARF max : půl (10762) ARF moyen : 1313	Nombre : 68 8.5 % de mots 5.9 % de ARF total ARF max : centrum (10120) ARF moyen : 995

a

mots affichés : 800 (sur 4809) | ARF moyen : 2612 | le plus fréquent : doba (61603) | le moins fréquent : konzerva (393)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 33 4.1 % de mots 2.8 % de ARF total ARF max : předseda (14214) ARF moyen : 1754	Nombre : 1 0.1 % de mots 0.1 % de ARF total ARF max : chleba (1612) ARF moyen : 1612	Nombre : 761 95.1 % de mots 96.6 % de ARF total ARF max : doba (61603) ARF moyen : 2653	Nombre : 5 0.6 % de mots 0.5 % de ARF total ARF max : téma (7208) ARF moyen : 2217

consonne molle

mots affichés : 800 (sur 1601) | ARF moyen : 913 | le plus fréquent : konec (31256) | le moins fréquent : harmonikář (51)

Le rapport des genres en fonctions du nombre des mots correspondants au motif ...

Le rapport des genres en fonction de la somme de l'ARF...

Le rapport des genres en fonction de l'ARF moyen...

Masculins animés	Masculins inanimés	Féminins	Neutres
Nombre : 335 41.9 % de mots 27.5 % de ARF total ARF max : muž (22138) ARF moyen : 601	Nombre : 244 30.5 % de mots 37 % de ARF total ARF max : konec (31256) ARF moyen : 1107	Nombre : 221 27.6 % de mots 35.5 % de ARF total ARF max : věc (26962) ARF moyen : 1172	Nombre : 0 0 % de mots 0 % de ARF total ARF max : (0) ARF moyen : 0

Déclinaison et genre grammatical en tchèque

- langue slave occidentale
- **flexion nominale** très **riche** : sept cas (nominatif, génitif, datif, accusatif, vocatif, locatif, instrumental), quatre genres (masculins animé et inanimé, féminin, neutre), deux nombres (singulier, pluriel) avec des résidus du duel.
- **nombreux paradigmes** de déclinaison avec des sous-types et des exceptions.
- des **variantes** (doublets, triplets) de registre et/ou de fonction.
- alternances **vocaliques** et **consonantiques** du radical lors de la déclinaison.
- **système vocalique** avec une opposition phonologique de longueur (quantité vocalique).
- **système consonantique** relativement complexe distinguant trois grandes classes de consonnes :
  - **consonnes dures** : h, ch, k, g, r, d, t, n
  - **consonnes molles** : c, č, ď, j, ň, ř, š, ť
  - **consonnes mixtes** : b, f, l, m, p, s, v, z
- grâce à la flexion, l'ordre des mots est géré par la structure informationnelle de la phrase.

Exemple d'un paradigme neutre

	singulier	pluriel
nominatif	*kuře*	*kuřata*
génitif	*kuřete*	*kuřat*
datif	*kuřeti*	*kuřatům*
accusatif	*kuře*	*kuřata*
vocatif	*kuře*	*kuřata*
locatif	*kuřeti*	*kuřatech*
instrumental	*kuřetem*	*kuřatům*

Exemple d'un paradigme féminin

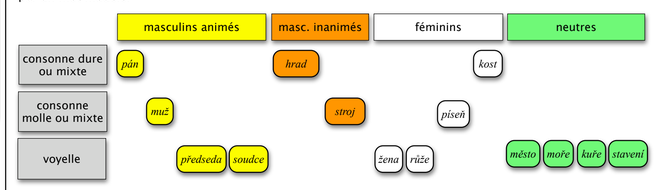
	singulier	pluriel
nominatif	*růže*	*růže*
génitif	*růže*	*růží*
datif	*růži*	*růžím*
accusatif	*růži*	*růže*
vocatif	*růže*	*růže*
locatif	*růži*	*růžích*
instrumental	*růží*	*růžemi*

Exemple d'un paradigme masculin

	singulier	pluriel
nominatif	*soudce*	*soudci*
génitif	*soudce*	*soudců*
datif	*soudci*	*soudcům*
accusatif	*soudce*	*soudce*
vocatif	*soudce*	*soudce*
locatif	*soudci*	*soudcích*
instrumental	*soudcem*	*soudci*

Paradigmes de la déclinaison nominale tchèque et le genre grammatical

Les grammaires traditionnelles reconnaissent 14 types de déclinaison nominale. Le type paradigmatique d'un substantif (l'ensemble de désinences s'attachant à son radical) est donné par son genre grammatical et par la nature morpho-phonologique de la périphérie droite (terminaison) de son lemme. Chaque type est représenté par un mot modèle.



Remarque : Au sein de chaque type, il existe de nombreux sous-types, définis par les différences par rapport au mot modèle. Cependant, ces différences touchent rarement la forme du lemme qui est déterminante pour l'identification du type paradigmatique (il existe notamment des glissements entre les types féminins *růže* et *píseň*) et il ne serait pas utile de les prendre en compte ici.

Erreur de genre sur CETLEF.fr

CETLEF.fr 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.** Erreur de genre est un des types d'une telle erreur.

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. Le diagnostic automatique des erreurs, implémenté sur cette plate-forme, est basé sur cette hypothèse. Il 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.

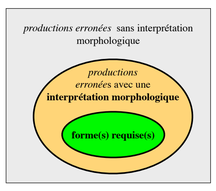
CETLEF.fr illustre les possibilités d'un riche modèle morphologique et des techniques de TAL employées dans un outil d'enseignement de langue assisté par ordinateur, voir par exemple (Heift & Schulze 2007).

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.

Définition d'une erreur de déclinaison

Comme erreur de déclinaison est considérée **toute production différente** de la (les) forme(s) requise(s) ayant une **interprétation morphologique.** Une **interprétation morphologique** peut être assignée à une production erronée, si elle correspond à une des formes hypothétiques générées à partir du radical de la forme requise par :

- emploi d'une **désinence inappropriée** appartenant (a) au paradigme de la forme requise, (b) appartenant à un paradigme différent
- simple **concaténation** du radical et d'une désinence **sans la réalisation** d'une **alternance obligatoire**
- erreurs d'ordre **graphique** : erreurs de diacritique, de casse



... plusieurs interprétations d'une forme erronée sont possibles, notamment à cause de l'homonymie des désinences.

Typologie des erreurs d'après l'attribut atteint et l'erreur de genre

La typologie formelle des erreurs de déclinaison est basée sur les différences entre les valeurs des attributs morphologiques de la forme requise et celles de la forme hypothétique correspondant à la production erronée.

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**

Exemple de l'interprétation d'une production erronée en tant qu'**erreur de genre :**

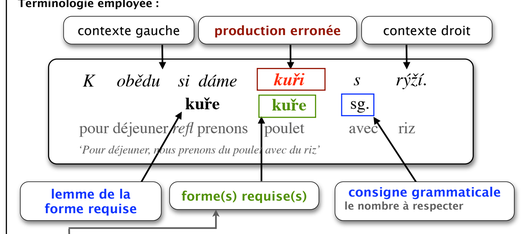
forme requise *kuře*	≠	production erronée *kuři*
cas : *acc*		*cas* : *acc*
num : *sg*		*num* : *sg*
gen : n		**gen : f**
alt : *sans*		*alt* : *sans*
pdgm : kr		**pdgm : rz**
tagMorph : *N*		*tagMorph* : *N*
dia : 1		*dia* : 1
casse : 1		*casse* : 1

Comme **erreur de genre** est donc considérée toute forme qui peut être interprétée comme ayant une désinence exprimant le cas et le nombre corrects, mais qui a ces fonctions au sein d'un paradigme de genre différent.

Tâche de déclinaison au sein d'un exercice à trous

Exemple : *«K obědu si dáme ..... (kuře, sg.) s rýží. »*

Terminologie employée :



Annotation morphologique de la forme requise

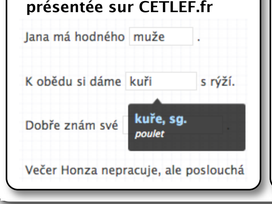
(1) lemme de la forme requise, (2) catégorie lexicale, (3) type morphologique
(4) paradigme, (5) cas, (6) genre, (7) nombre, (8) alternance éventuelle

Exemple

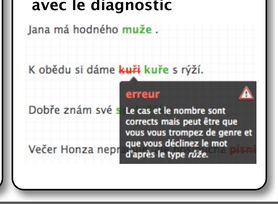
requis	lemme	tagLex	tagMorph	pdgm	cas	num	gen	alt
kuře	kuře	subst	N	kr	acc	sg	n	sans

Remarque : L'annotation est assignée automatiquement à la forme requise à l'aide d'un générateur morphologique rudimentaire et vérifiée par l'auteur des exercices.

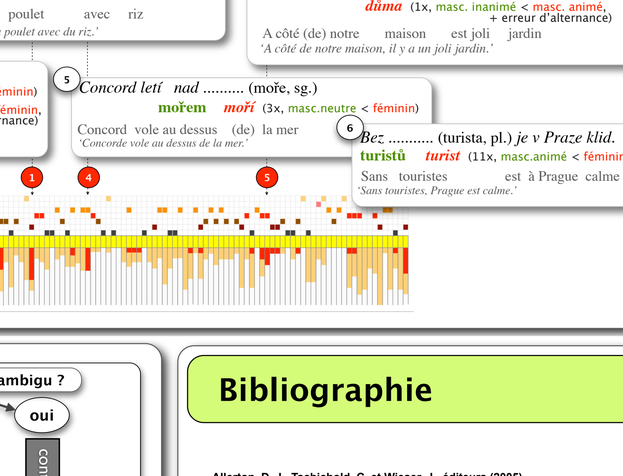
Une tâche de déclinaison présentée sur CETLEF.fr



La correction de la tâche avec le diagnostic



Diagnostic automatique des erreurs



Exemple :

kuře
kuře|subst|N|kr|acc|sg|n|sans
kuři

Diagnostic

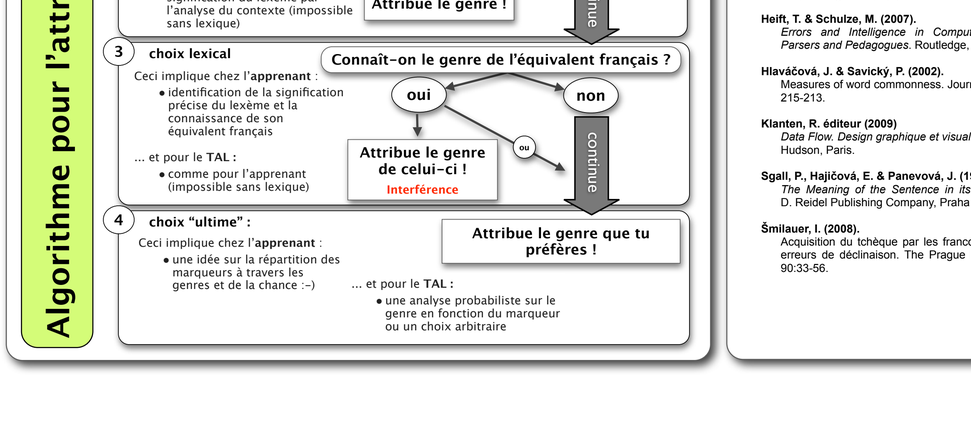
```
HOR EXT|rz|i|acc|sg|f|sans|erreur_gen
```

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 růže.

Entre 85 et 90 % de productions collectées au sein de CETLEF sont interprétées comme une erreur de déclinaison.

Erreurs de genre dans les productions recueillies

Le graphique ci-dessous présente les erreurs de genres commises par les apprenants dans 162 tâches de déclinaison au sein de 16 exercices sur CETLEF.fr. A l'aide de cette affichage, possible grâce à la base de donnée et au diagnostic des erreurs, il est aisé d'observer les différentes occurrences des erreurs de genre et leur relation avec la nature du marqueur formel du lemme de la production requise. Avec plus de données recueillies dans le futur, il sera possible de mieux dégager des tendances générales.



Quelques exemples des erreurs de genre :

1 *Na ..... (pánev, sg.) osmahneme dvě cibule.*
pánvi ***pánvu*** (4x, féminin < masc.inanimé) ***pánvě*** (5x, féminin < masc.inanimé)
pánve (1x, féminin < masc.inanimé + erreur de diacritique)
Sur poêle (faisons roussir) deux oignons
'Nous faisons roussir deux oignons dans la poêle.'

2 *K obědu si dáme .....* (kuře, sg.) *s rýží.*
kuře ***kuři*** (7x, neutre < féminin)
pour déjeuner *refl* prenons poulet avec riz
'Pour déjeuner, nous prenons du poulet avec du riz.'

3 *Vedle našeho .....* (dům, sg.) *je hezká zahrada.*
domu ***doma*** (15x, masc. inanimé < masc. animé)
důma (1x, masc. inanimé < masc. animé, + erreur d'alternance)
A côté (de) notre maison est joli jardin
'A côté de notre maison, il y a un joli jardin.'

4 *Hanička se líbí .....* (Jirka, sg.).
Jirkovi ***Jirce*** (6x masc.animé < féminin)
Jirkě (1x, masc.animé < féminin, + erreur d'alternance)
Hanička *refl* plaît (à) Jirka
'Hanička plaît à Jirka.'

5 *Concord letí nad .....* (moře, sg.)
mořem ***moří*** (3x, masc.neutre < féminin)
Concord vole au dessus (de) la mer
'Concorde vole au dessus de la mer.'

6 *Bez .....* (turista, pl.) *je v Praze klid.*
turistů ***turist*** (11x, masc.animé < féminin)
Sans touristes est à Prague calme
'Sans touristes, Prague est calme.'

Algorithme pour l'attribution du genre

1 choix formel
Ceci implique chez l'**apprenant** :
- connaissance des marqueurs et de leurs propriétés

... et pour le **TAL :**
- représentation formelle des propriétés des marqueurs

Le marqueur est-il ambigu ? → non → Attribue le genre ! ; oui → continue

2 choix sémantique
Ceci implique chez l'**apprenant** :
- identification approximative de la signification du lexème

... et pour le **TAL :**
- identification approximative de la signification du lexème par l'analyse du contexte (impossible sans lexique)

Peut-on considérer le genre naturel ? → oui → Attribue le genre ! ; non → continue

3 choix lexical
Ceci implique chez l'**apprenant** :
- identification de la signification précise du lexème et la connaissance de son équivalent français

... et pour le **TAL :**
- comme pour l'apprenant (impossible sans lexique)

Connaît-on le genre de l'équivalent français ? → oui → Attribue le genre de celui-ci ! (Interférence) ; non → ou → continue

4 choix "ultime" :
Ceci implique chez l'**apprenant** :
- une idée sur la répartition des marqueurs à travers les genres et de la chance :–)

... et pour le **TAL :**
- une analyse probabiliste sur le genre en fonction du marqueur ou un choix arbitraire

Attribue le genre que tu préfères !

Bibliographie

Allerton, D. J., Tschichold, C. et Wieser, J., éditeurs (2005).
Linguistics, Language Learning and Language Teaching. Schwabe, Basel.

Čermák, F. & Křen, M. (2004).
Frekvenční slovník češtiny. Nakladatelství Lidové Noviny, Praha.

Čermák, F. & Křen, M. (2005).
New generation corpus-based frequency dictionaries: The case of czech. International Journal of Corpus Linguistics, 10:453-467.

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

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

Hlaváčová, J. & Savický, P. (2002).
Measures of word commonness. Journal of Quantitative Linguistics, 9(3): 215-213.

Klanten, R. éditeur (2009)
Data Flow. Design graphique et visualisation d'information. Thames & Hudson, Paris.

Sgall, P., Hajičová, E. & Panevová, J. (1986).
The Meaning of the Sentence in its Semantic and Pragmatic Aspects. D. Reidel Publishing Company, Praha - Amsterdam.

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