*RÉCITAL 2004, Fès, 21 avril 2004*

# Reconnaissance automatique des adjectifs durs et des adverbes réguliers lors de l'analyse morphologique automatique du slovaque

**Diana Jamborova-Lemay**

**CERTAL – INALCO**
**73, rue Broca, 75013 Paris**
**dianalemay@aol.com**

## Résumé – Abstract

L'analyse morphologique automatique du slovaque constitue la première étape d'un système d'analyse automatique du contenu des textes scientifiques et techniques slovaques. Un tel système pourrait être utilisé par des applications telles que l'indexation automatique des textes, la recherche automatique de la terminologie ou par un système de traduction. Une description des régularités de la langue par un ensemble de règles ainsi que l'utilisation de tous les éléments au niveau de la forme du mot qui rendent possible son interprétation permettent de réduire d'une manière considérable le volume des dictionnaires. Notamment s'il s'agît d'une langue à flexion très riche, comme le slovaque. La reconnaissance automatique des adjectifs durs et des adverbes réguliers constitue la partie la plus importante de nos travaux. Les résultats que nous obtenons lors de l'analyse morphologique confirment la faisabilité et la grande fiabilité d'une analyse morphologique basée sur la reconnaissance des formes et ceci pour toutes les catégories lexicales.

Automatic morphological analysis of Slovak language is the first level of an automatic analyser for Slovak scientific and technical texts. Such a system could be used for different applications: automatic text indexation, automatic research of terminology or translation systems. Rule-based descriptions of language regularities as well as the use of all the formal level elements of words allow reducing considerably the volume of dictionaries. Notably in case of inflectionally rich languages such as Slovak. The most important part of our research is the automatic recognition of adjectives and regular adverbs. The results obtained by our morphological analyser justify such an approach and confirm the high reliability of morphological analysis based on form-recognition for all lexical categories.

## Mots-clefs – Keywords

Traitement automatique du slovaque, Analyse morphologique, Morphologie flexionnelle, Morphologie dérivationnelle.

Natural Language Processing of Slovak, Morphological Analysis, Inflectional Morphology, Derivational Morphology.

# 1 Analyse morphologique

Nos travaux en analyse morphologique automatique du slovaque s'appuient sur les travaux déjà réalisés au sein du CERTAL et notamment sur ceux de Patrice Pognan[1] (CERTAL – INALCO) en tchèque. L'analyseur morphologique du slovaque fonctionne en parallèle avec celui de Pognan pour le tchèque. Le texte d'entrée peut donc être écrit soit en tchèque soit en slovaque. Le premier module qui détermine la langue d'entrée est commun aux deux analyseurs. Ce fonctionnement justifie donc notre décision d'adopter les mêmes approches en morphologie que celles de Pognan et ceci d'autant plus que les deux langues présentent les mêmes caractéristiques morphologiques. L'analyseur morphologique du slovaque prouve que les approches de Pognan pour le tchèque sont aussi valables pour d'autres langues proches et notamment pour le slovaque.

L'analyse morphologique qui a pour but d'extraire tous les éléments d'une correspondance forme-fonction servira, appuyée par des procédés contextuels, de point de départ à l'analyse syntaxique du slovaque. L'essentiel de notre travail consiste à étudier attentivement la suffixation slovaque et d'exploiter les possibilités qu'offre la morphologie de cette langue pour un traitement automatique. Le slovaque étant une langue à flexion avec un très riche système de dérivation, l'utilisation exclusive des dictionnaires et des méthodes statistiques ne nous semble pas être la mieux adaptée.

# 2 Fonctionnement de l'analyseur

L'analyseur morphologique du slovaque est un automate[2]. Il fonctionne par étapes. Le corpus est tout d'abord découpé en mots, ensuite l'analyse fonctionne comme un système d'apprentissage qui nous permet de constituer un lexique des mots du texte analysé. L'analyseur est implémenté en C++. Le lexique est gardé dans le conteneur associatif de type « map », où la clé est un mot du corpus et la valeur associée est une structure qui contient toutes les informations que l'analyseur obtient au cours de l'analyse.

---

[1] Pognan P. (1992), *Automatické zpracování češtiny pro vědecko-technické informace*, 16th World Congress of SVU, Prague, Společnost pro Vědu a Umění.

Pognan P. (1996), *Approches grammaticale et textuelle pour l'élaboration de système d'analyse automatique et d'indexation terminologique*, Journées Realiter, Nice.

[2] Jamborova-Lemay D. (2003), *Analyse morphologique automatique du slovaque.* Thèse de doctorat présentée à l'INALCO, Paris.

# 3 Reconnaissance automatique des adjectifs durs

La catégorie de l'adjectif est probablement la catégorie lexicale la moins ambiguë pour la reconnaissance automatique du slovaque. Les adjectifs durs[3] n'ont qu'un seul modèle de déclinaison " pekný " à la différence des substantifs, où le slovaque connaît treize modèles. Le modèle de déclinaison pour les adjectifs durs "pekný" (*joli*) régit aussi la déclinaison de certains pronoms et adjectifs numéraux et des participes passés passifs. Les pronoms constituent une catégorie lexicale fermée et leur nombre est de quelques dizaines. Ils sont stockés dans une liste de mots outils, ce qui permet leur identification par simple consultation de cette liste. En ce qui concerne les adjectifs numéraux et les participes passés nous avons préféré dans un premier temps ne pas faire la distinction entre adjectifs durs propres, adjectifs numéraux, adjectifs issus du participe passé passif et participes passés passifs.

Les adjectifs durs slovaques présentent une particularité par rapport aux adjectifs durs tchèques. La loi de rythme, qui est une particularité du slovaque interdit la succession de deux syllabes longues au sein d'un mot et complique ainsi la reconnaissance des désinences univoques en slovaque. L'application de cette loi a pour conséquence l'apparition de désinences raccourcies qui perdent ainsi leur caractère univoque (le « -y » bref de « krátky » ne permet pas à lui seul d'attribuer la catégorie lexicale).

## 3.1 Reconnaissance des désinences univoques

La reconnaissance automatique des adjectifs durs s'appuie tout d'abord sur la reconnaissance des désinences univoques : "-ý", "-ého", "-ých", "-ými". Nous utilisons la reconnaissance des désinences adjectivales univoques pour créer une liste dynamique de radicaux adjectivaux. Ceci nous permet d'élargir la reconnaissance des adjectifs à tous les adjectifs qui ont au moins une occurrence à désinence univoque dans le corpus. Dans ces cas nous conservons le radical dans une liste et procédons ensuite à la reconnaissance du couple "radical adjectival - désinence adjectivale".

| **Désinence** | **Cat. Lex.** | **Genre** | **Nombre** | **Cas** |
|---|---|---|---|---|
| -ý | Adjectif dur | Masculin animé | Singulier | Nominatif |
| | | Masculin inanimé | Singulier | Nominatif |
| | | | | Accusatif |
| -ého | Adjectif dur | Masculin animé | Singulier | Accusatif |
| | | Masculin inanimé | Singulier | Génitif |
| | | Neutre | Singulier | Génitif |

Tableau 1. Valeur des désinences adjectivales univoques « -ý » et « -ého »

[3] Est considéré comme adjectif dur tout adjectif qualificatif terminé au nominatif singulier en « -ý/y », voyelle dure par opposition à « i », voyelle molle. Le système d'opposition dure - molle, valable pour les voyelles et les consonnes, sous-tend l'ensemble des paradigmes et des alternances consonantiques et certaines alternances vocaliques.

### 3.2 Reconnaissance du segment suffixe-désinence

Lorsque la désinence est ambiguë nous procédons à la reconnaissance de quelques segments "suffixe-désinence". Alors que la reconnaissance d'un adjectif dur terminé par une désinence univoque est simple, elle se complique pour le même adjectif dont la désinence a la même forme qu'une désinence d'une autre catégorie lexicale. Dans ces cas il peut être utile d'avoir recours à la reconnaissance du suffixe qui précède la désinence. Nous procédons à une étude approfondie de quelques suffixes parmi les plus fréquents. Il s'agit de suffixes liés aux variantes longues des désinences adjectivales: " -ský " , " -cký " , " -avý " , " -ivý " , " -ový ".

| **Suffixe** | **Désinence** | **Cat. Lex.** | **Genre** | **Nombre** | **Cas** |
|---|---|---|---|---|---|
| -sk-<br>-ck-<br>-av-<br>-iv-<br>-ov- | -om | Adjectif dur | Masculin | Singulier | Locatif |
| | | | Neutre | Singulier | Locatif |
| | | Substantif[4] | Masculin | Singulier | Instrumental |
| | | | Neutre | Singulier | Instrumental |

Tableau 2. Valeurs des segments " suffixe adjectival – désinence –om "

Prenons comme exemple la désinence " -om ". Seule, elle ne suffit pas à indiquer les valeurs morphologiques. C'est une désinence très ambiguë. En élargissant la reconnaissance au segment "suffixe adjectival – désinence " nous avons la possibilité d'obtenir les valeurs morphologiques univoques comme le montre le tableau 2.

Notons aussi qu'un grand nombre d'adjectifs d'origine étrangère est reconnu lors de la procédure de reconnaissance des mots d'emprunt.

## 4 Reconnaissance dynamique des adverbes

Les adverbes sont pour la plupart formés à partir des adjectifs qualificatifs à l'aide de trois suffixes: "-o", "-e", "-y". Certains adverbes sont à l'origine des substantifs (avec ou sans préposition). Ils se sont figés et devenus adverbes. D'autres encore ont été formés à partir de formes verbales, souvent à l'aide du suffixe "-mo". Seuls les adverbes formés à partir des adjectifs de manière régulière sont reconnaissables automatiquement. Pour les autres adverbes, nous avons une liste non exhaustive d'adverbes irréguliers les plus utilisés. Nous envisageons aussi une possibilité de reconnaissance de certains adverbes au niveau de l'analyse syntaxique. Nous avons sélectionné certains segments univoques ( "-úco", "-avo", "-ovo", "-ovane", "-cky", "-sky") qui permettent une reconnaissance d'adverbes rapide et simple.

[4] Le nombre de substantifs étant très réduit, nous les considérons comme exceptions et ils sont testés préalablement.

Dans la deuxième phase nous avons recours aux adjectifs durs déjà reconnus. Lors de la procédure de reconnaissance des adjectifs, les radicaux d'adjectifs reconnus à l'aide des désinences univoques sont stockés dans une liste dynamique tout au long de l'analyse. Nous utilisons cette même liste dans la reconnaissance dynamique des adverbes réguliers. Si nous rencontrons une forme lexicale terminée en "-o" ou en "-e", suffixes formateurs d'adverbes, nous la comparons avec la liste de radicaux adjectivaux. Ce procédé nous permet de reconnaître tous les adverbes réguliers formés à partir d'adjectifs déjà reconnus dans le texte analysé.

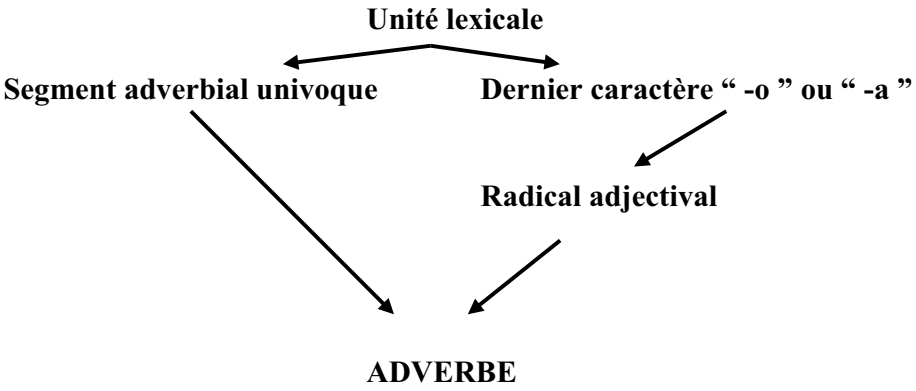


Figure 3 : Schéma de reconnaissance des adverbes réguliers

A la sortie de l'analyse morphologique tous les adverbes réguliers du corpus traité qui sont dérivés à partir d'adjectifs durs reconnus préalablement sont identifiés.

# 5 Evaluation des résultats

Notre corpus a été réalisé à partir de textes slovaques scientifiques et techniques issus de la revue technique bilingue tchèque/slovaque « Mechanizace zemědělství » ( tous les n° de 1996) et il contient près de 22 000 formes. Le lexique obtenu par l'analyseur en contient 6456. (Nous ne conservons qu'une seule occurrence de chaque forme de mot du corpus.)Notre évaluation est réalisée sur ce lexique tassé. Les résultats sont calculés par rapport à l'étiquetage manuel que nous avons effectué sur la totalité du lexique.

Le taux d'erreurs très faible (0,07%) montre que le système de reconnaissance est très fiable et que nous pourrons nous appuyer sur les résultats de l'analyse morphologique dans les étapes suivantes de l'analyse.

La catégorie des adjectifs est la catégorie la mieux reconnue. Le pourcentage des adjectifs durs reconnus s'élève à 78,5%. Nous avons également pris en compte les adjectifs qui ont été reconnus en double catégorisation. La majorité des adjectifs durs non reconnus sont les adjectifs issus du participe passé passif terminés en -ný, -tý qui n'ont pas de forme avec désinence univoque dans le corpus, par exemple:

| | |
|---|---|
| hodnotné | *précieux* |
| novovyvinuté | *nouvellement développé* |

Pour l'instant ce type d'adjectifs ne peut être reconnu que s'il existe au moins une occurrence avec désinence univoque dans le corpus. Plus le corpus est important plus la probabilité d'y trouver une forme à désinence univoque est grande. Ce qui veut dire que ce système de reconnaissance est encore plus efficace pour les grands corpus. Néanmoins nous envisageons de continuer et d'approfondir la reconnaissance de la suffixation des adjectifs durs issus du participe passé passif afin d'améliorer le taux de reconnaissance des adjectifs durs.

Les adverbes non reconnus (47%) sont en majorité des adverbes irréguliers. Nous envisageons d'élargir notre dictionnaire des adverbes irréguliers afin d'obtenir une reconnaissance plus complète des adverbes irréguliers. Mais on retrouve aussi une partie des adverbes réguliers formés à partir d'adjectifs durs (20%). Les adverbes réguliers non reconnus sont soit formés à partir d'adjectifs qui ne se trouvent pas dans le corpus ("exaktne" *exactement)* soit qui n'ont pas été reconnus ("jednoducho" *simplement*). Cependant nous constatons que plus de la moitié (53%) des adverbes du lexique a été reconnue. La reconnaissance d'adverbes sera complétée au niveau d'étapes ultérieures, en particulier syntaxique.

# 6 Conclusion

Nous savons que des travaux importants qui concernent l'élaboration du Corpus National slovaque ont été entrepris en Slovaquie au sein de l'Académie des Sciences.
Le projet de création du Corpus national de la langue slovaque a été approuvé par le gouvernement de la République Slovaque en février 2002. Son objectif est de former une structure de travail qui permettrait la création du Corpus National de la langue slovaque et de s'en donner les moyens matériels et techniques. Le Corpus doit être une institution nationale accessible au large public par l'intermédiaire d'Internet. La première partie du Corpus National est accessible au public depuis septembre 2003, mais les slovaques manquent encore de nombreux outils pour le traitement automatique du slovaque.
A notre connaissance il n'existe pas d'analyseur morphologique du slovaque comparable au nôtre. L'utilisation d'un analyseur morphologique automatique pourrait apporter une aide précieuse à nos collègues slovaques.

# Références

Bujalka A., Dubníček J. (1998), *Slovenský jazyk II - Morfológia*, Bratislava, Univerzita Komenského.

Kačala J., Pisárčiková M. (1997)**,** *Krátky slovník slovenského jazyka*, Bratislava, Veda.

Mistrík J. (1983), *Moderná slovenčina*, Bratislava, SPN.

Mistrík J. (1976), *Retrográdny slovník slovenčiny*, Bratislava, Univerzita Komenského.

Oravec J., Bajzíková E., Furdík J. (1984), *Morfológia, Súčasný spisovný jazyk*, Bratislava, SPN.

Pauliny E. (1997), *Krátka gramatika slovenská*, Bratislava, Národné literárne centrum.

Pognan P. (1996), *Approches grammaticale et textuelle pour l'élaboration de système d'analyse automatique et d'indexation terminologique*, Journées Realiter, Nice.

Pognan P. (1998), *Analyse automatique du tchèque sur la base de l'étude de la grammaire et du texte. Quelle stratégie syntaxique choisir?* in:Issues of Valency and Meaning, Studies in Honour of Jarmila Panevová, Prague, Karolinum,Charles University Press.