UNIVERSITÉ DE GRENOBLE - ALPES

UFR - LANGAGE, LETTRES ET ARTS DU SPECTACLE, INFORMATION ET COMMUNICATION

DOCUMENTATION, ENSEIGNEMENT ET TRAITEMENT DES LANGUES PEU-DOTÉES : SUR UN EXEMPLE DES LANGUES SLAVES DU SUD – LE MACÉDONIEN

JOVAN KOSTOV

E.A. 4514 - PLIDAM

PARCOURS

- 2000 - Baccalauréat scientifique (option maths) - Gevgelija, Macédoine
- 2002 - Diplôme universitaire de premier degré en langues (français / espagnol) - Université de Skopje, Macédoine
- 2005 - Licence de Sciences du langage TAL - Montpellier 3
- 2007 - Master de Sciences du langage TAL - Toulouse / Montpellier (travail sur les genres du discours)
- 2013 - Doctorat TAL - INALCO (travail sur le système verbal du macédonien : génération automatique des formes verbales)
- 2013 - 2015 - ATER - Université Paris - Sorbonne (informatique : langues et sciences humaines; TAL)
- Septembre 2016 - Post-doc - INALCO (valorisation des ressources et méthodes de langues).

QU'ALLONS-NOUS DISCUTER?

- Que signifie le terme « langue peu-dotée »?
- Pourquoi travailler sur les langues peu-dotées?
- Quelques pistes pour travailler sur les langues peu-dotées :
  - Approche dite « des langues voisines »
  - Utilité des réseaux sociaux dans la construction des ressources pour les langues peu-dotées.
- Quelques exemples :
  - FlexiMac 1.1. - conjugueur de verbes macédoniens
  - Pangloss - documentation des langues « en danger » ou des langues peu-dotées à l'échelle mondiale.

LANGUES PEU-DOTÉES?

« faible diversité linguistique en TAL »
(Enguehard & Mangeot, 2014)

Constat:

- 7200 articles publiés entre 2000 et 2014
- LREC & ACL

	anglais	alle-mand	fran-çais	chinois	espa-gnol	japo-nais
ACL & LREC	64%	21%	20%	17%	15%	12%
ACL	62%	14%	12%	23%	9%	12%
LREC	65%	25%	25%	13%	19%	12%

Table 1 : Les six langues très bien dotées (mentionnées dans plus de 10% des articles)

LANGUES PEU-DOTÉES?

Autres exemples :

- Langues à tradition orale.

- Langues peu-utilisées (quelques exemples : variétés du valaque dans les Balkans au profit des langues « officielles », le pomak, les langues indigènes du Mexique au profit de l'espagnol).

- Langues peu décrites (surtout les langues récemment standardisées).

LANGUES PEU-DOTÉES : QUELLES IMPLICATIONS POUR L'INFORMATICIEN-LINGUISTE?

LANGUE BIEN-DOTÉE = BOULEVARD

LANGUE PEU-DOTÉE = IMPASSE

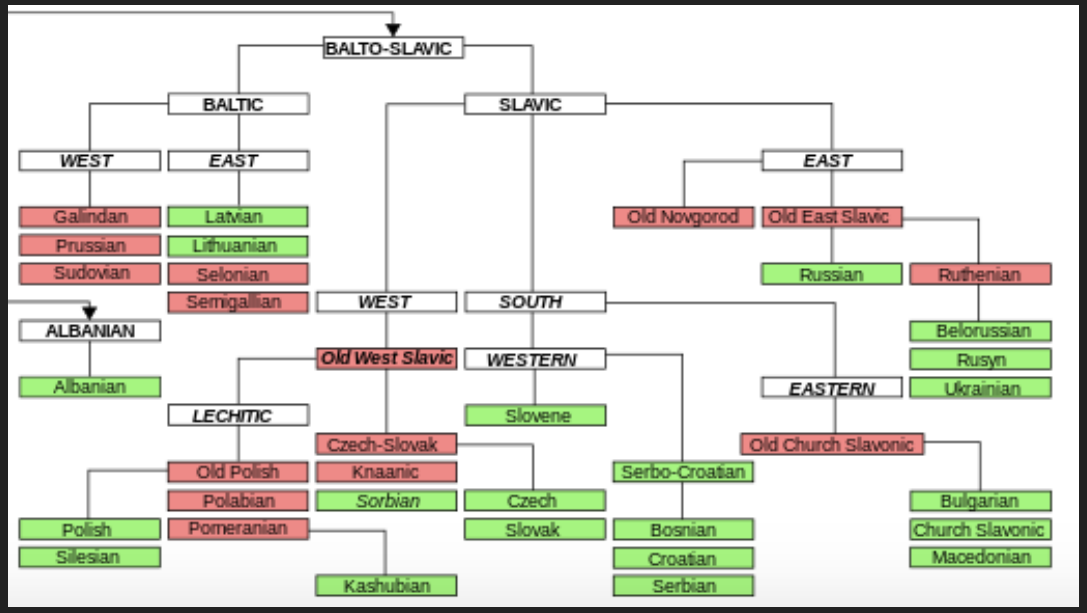
LANGUES PEU-DOTÉES : QUELLES IMPLICATIONS POUR L'INFORMATICIEN-LINGUISTE?

Le macédonien?

Langue slave du Sud - standardisation en 1945s

Environ 2 millions de locuteurs

9 dialectes (« parlers ») principalement en Macédoine mais aussi au Nord de la Grèce et en Bulgarie



LANGUES PEU-DOTÉES : QUELLES IMPLICATIONS POUR L'INFORMATICIEN-LINGUISTE?

TAL et descriptions linguistiques

« Le verbe macédonien : pour un traitement automatique de nature linguistique et applications didactiques - Réalisation d'un conjugueur »

ETAPES:

- Analyse du comportement morphologique des verbes
- Réorganisation par classe flexionnelle (en fonction du comportement du thème)
- Modélisation informatique des verbes
- Intégration des algorithmes dans une plateforme qui « calcule » automatiquement et organisation en fonction des catégories grammaticales (mode, temps, personne)

- Statistiques :

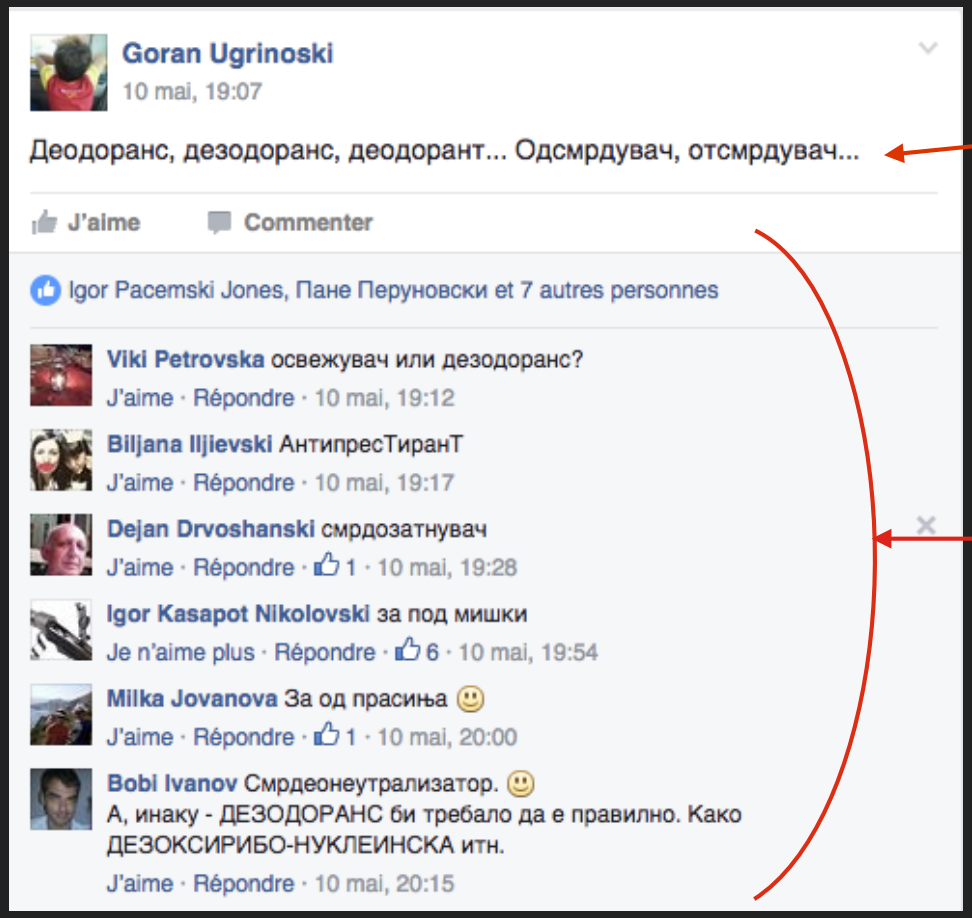
  Environ 25000 verbes testés
  91 % de conjugaisons correctes
  9 % irréguliers et défectueux

Résultat : http://fleximac.free.fr

LANGUES PEU-DOTÉES : QUELLES IMPLICATIONS POUR L'INFORMATICIEN-LINGUISTE?

Et comment fait-on fait quand on ne sait pas? Et quand on hésite?

Utilisation des Réseaux sociaux (exemple des groupes de Facebook)



IDENTIFICATION DU PROBLÈME

PROBLÈME DE CHOIX D'UNE UNITÉ LEXICALE (DÉODORANT) :

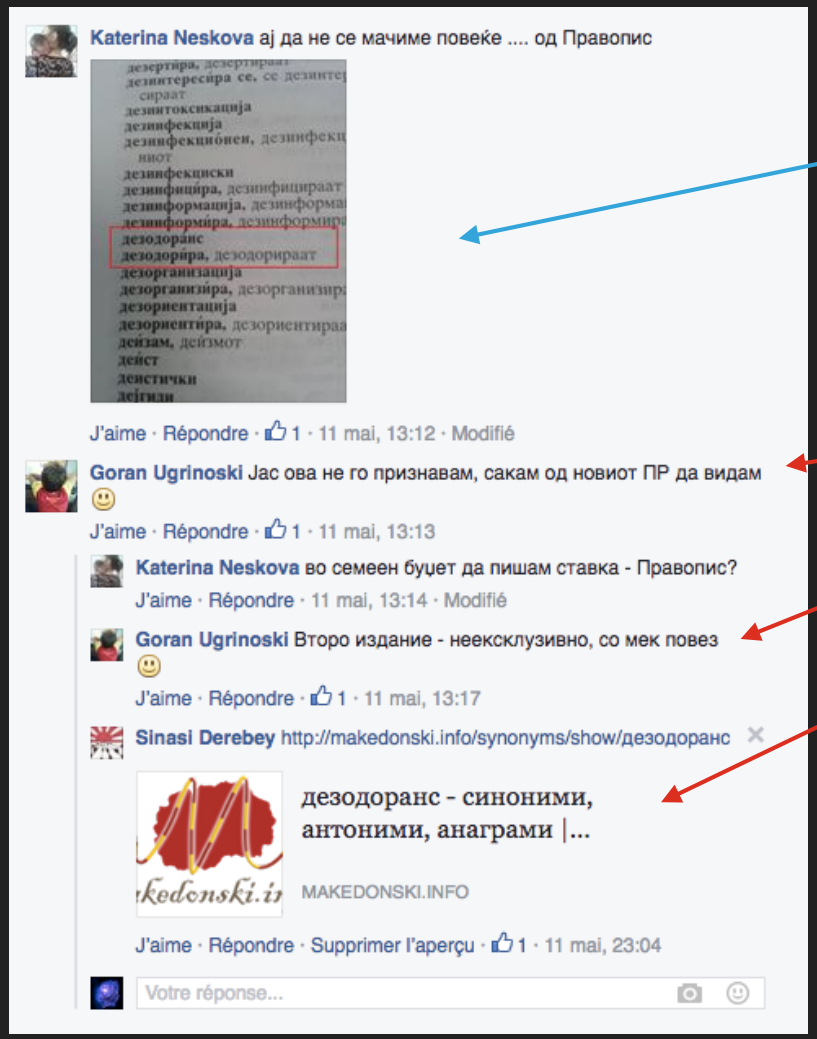
PLUSIEURS SOLUTIOINS:
- SANS LIAISON (/DEODORANS/)
- AVEC LIAISON (/DEZODORANS/)
- AVEC UNE TERMINAISON DIFFÉRENTE QUE CELLE D'USAGE (/DEODORANT/)

SOLUTIONS ALTERNATIVES:
- OTSMRDUVAC (TRADUCTION APPROXIMATIVE: [CE] QUI NEUTRALISE L'ODEUR) - LE PRÉFIXE OT- EST UN PRÉFIXE TRÈS PRODUCTIF POUR CRÉER DES NOMS D'AGENTS.

LANGUES PEU-DOTÉES : QUELLES IMPLICATIONS POUR L'INFORMATICIEN-LINGUISTE?

Apprentissage « entre pairs »

Utilisation des Réseaux sociaux (exemple des groupes de Facebook)



SOLUTION À PARTIR D'UNE RÉFÉRENCE (DICTIONNAIRE ORTHOGRAPHIQUE)

DISCUSSION SUR LA VALIDITÉ DE LA SOURCE :

- ELLE DATE DE 1991
- ON A BESOIN D'UNE SOURCE PLUS RÉCENTE (NOUVEAU DICTIONNAIRE ORTHOGRAPHIQUE PARU EN 2015)
- APPEL À UN DICTIONNAIRE ÉLECTRONIQUE QUI GÉNÈRE LES DÉFINITIONS ET L'ORTHOGRAPHE DES DIFFÉRENTS DICTIONNAIRES EXISTANTS LAISSANT LE CHOIX D'USAGE D'UN MOT DANS UNE FORME X OU Y (CERTAINE SOUPLESSE).

Atouts de l'apprentissage entre pairs : valorisation des connaissances

Le processus de formation est un processus réflexif : « doing is thinking of what we do » (D. Schönn)

ET LES AUTRES LANGUES?

Langues peu-dotées et/ou en extinction et/ou éteintes…

Projet Pangloss (resp. Alexandre François, UMR 7107 - LACITO)

- Description et documentation des langues du monde
- 140 langues
- 400 heures d'enregistrements (et leurs transcriptions et gloses)
- Ouverture vers le monde scientifique

VOIR SITE : http://lacito.vjf.cnrs.fr/pangloss/

QUELQUES RÉFÉRENCES INCONTOURNABLES

AUSTIN Peter & SALLABANK Julia (éd.) (2012), *The Cambridge Handbook of Endangered Languages*, Cambridge University Press, Cambridge.

Thèse de Vincent Berment (disponible sur http://portal.unesco.org/ci/fr/files/16735/10914394223these_Berment.pdf/these_Berment.pdf).

ENGUEHARD Chantal & MANGEOT Mathieu (2012), Favorisons la diversité linguistique en TAL (Journée d'étude de l'ATALA. "Ethique et Traitement Automatique des Langues", Nov 2014). Article disponible en ligne sur https://hal.archives-ouvertes.fr/hal-01096592/document).

Atelier « TAL et langues peu-dotées » (actes en ligne sur https://taln.limsi.fr/actes-articles.htm#tlpd).

Collection Pangloss : http://lacito.vjf.cnrs.fr/pangloss/

FlexiMac 1.1. : description détaillée sur https://www.academia.edu/26894236/FlexiMac_1.1._-_Conjugueur_automatique_des_verbes_mac%C3%A9doniens

Merci

Благодарам

Σας ευχαριστώ

Gracias

jovan.kostov@gmail.com - Jovan Kostov (PLIDAM, INALCO)