Transferts linguistiques des langues régionales vers le français, transferts technologiques du français vers les langues régionales

Ph. Boula de Mareüil (LIMSI, U. Paris Sud)

& D. Bernhard (LiLPa, U. Strasbourg)

Variation linguistique et Crowdsourcing :
Étudier la variation au 21ème siècle

23 octobre 2015

RESTAURE

RESSOURCES INFORMATISÉES ET TRAITEMENT AUTOMATIQUE POUR LES LANGUES RÉGIONALES

- PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Le projet RESTAURE

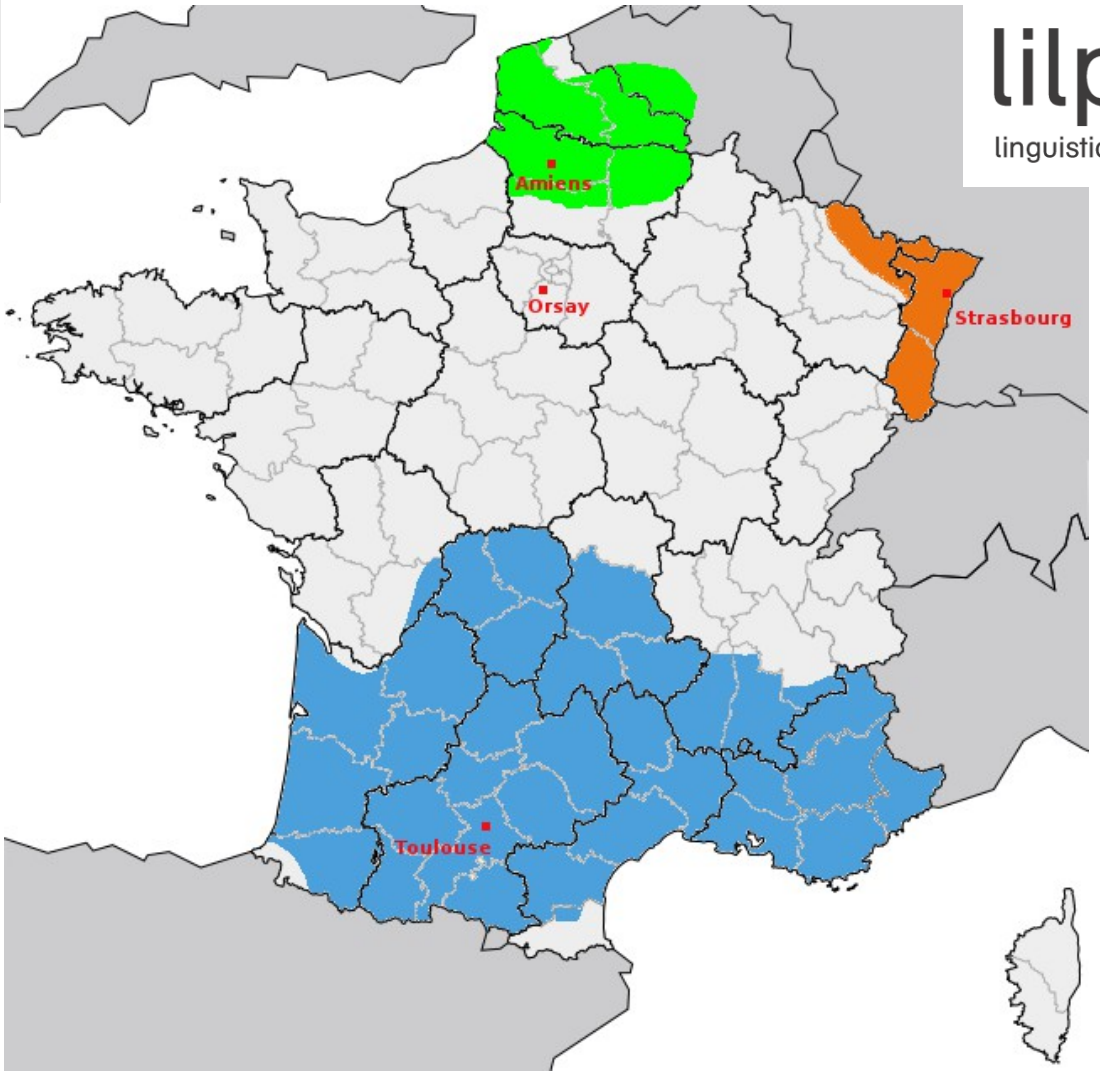
RESTAURE
RESSOURCES INFORMATISÉES ET TRAITEMENT AUTOMATIQUE POUR LES LANGUES RÉGIONALES

- Projet financé par l'ANR
- Début : janvier 2015
- Durée : 42 mois
- Objectifs :
  - Acquisition et normalisation de ressources (corpus et lexiques)
  - Développement d'outils pour l'acquisition et l'analyse de corpus
  - Diffusion auprès du grand public
- Site web : http://restaure.unistra.fr/

Partenaires et langues régionales

LESCLAP
linguistique et sociolinguistique :
contacts, lexique, appropriations, politiques

Picard
Resp : C. Rey



lilpa
linguistique, langues, parole

Alsacien
Resp : D. Bernhard

Clle ERSS

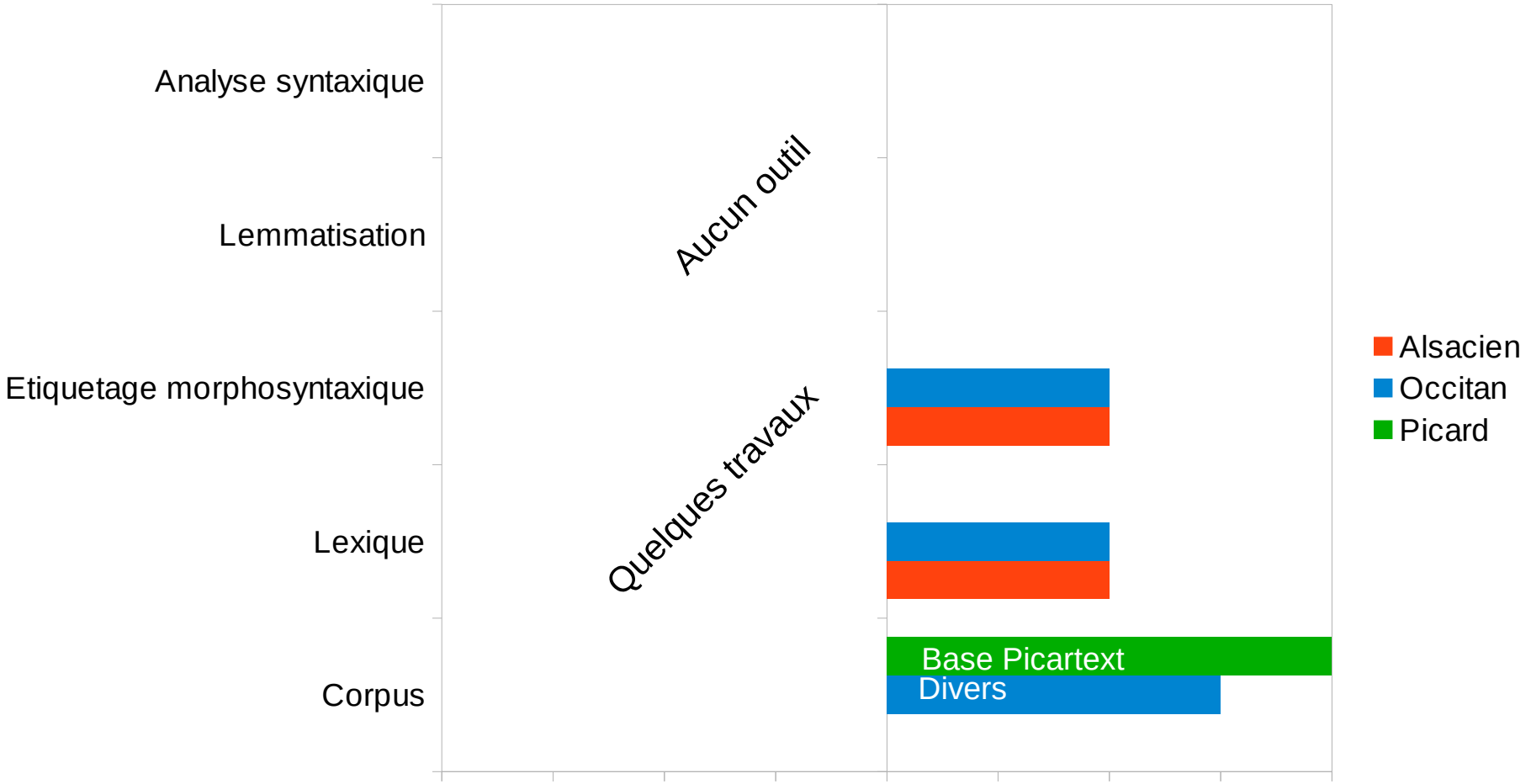
Occitan
Resp : M. Bras

Limsi

TAL
Resp. : A-L Ligozat

Motivations

Les langues régionales de France sont des langues peu dotées



Transferts technologiques depuis des langues "mieux dotées"

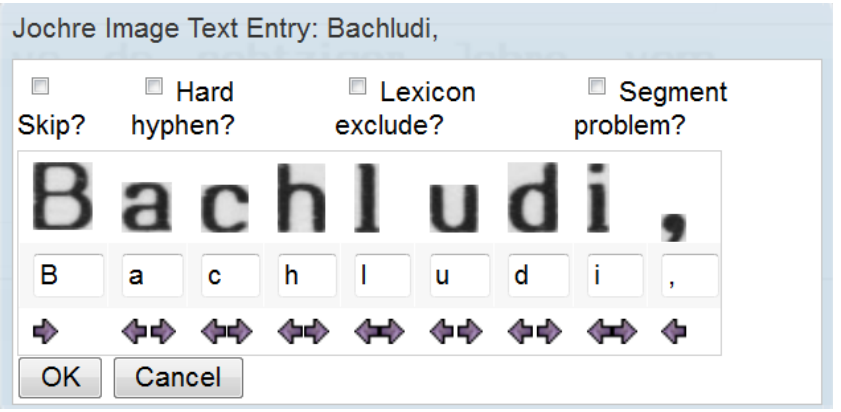
- Méthode : exploiter des outils et ressources disponibles pour des langues proches
  - Français pour le picard ?
  - Allemand pour l'alsacien ?
  - Catalan pour l'occitan ?

- Influence de l'application sur le choix de la langue "mieux dotée"
  - Exemple de la reconnaissance optique de caractères (OCR)

Principe de fonctionnement de l'OCR : exemple de Jochre
(Assaf Urieli, CLLE-ERSS Toulouse et Joliciel informatique)

Segmentation des images en paragraphes, lignes, mots et « formes »

Mise en correspondance forme - caractère



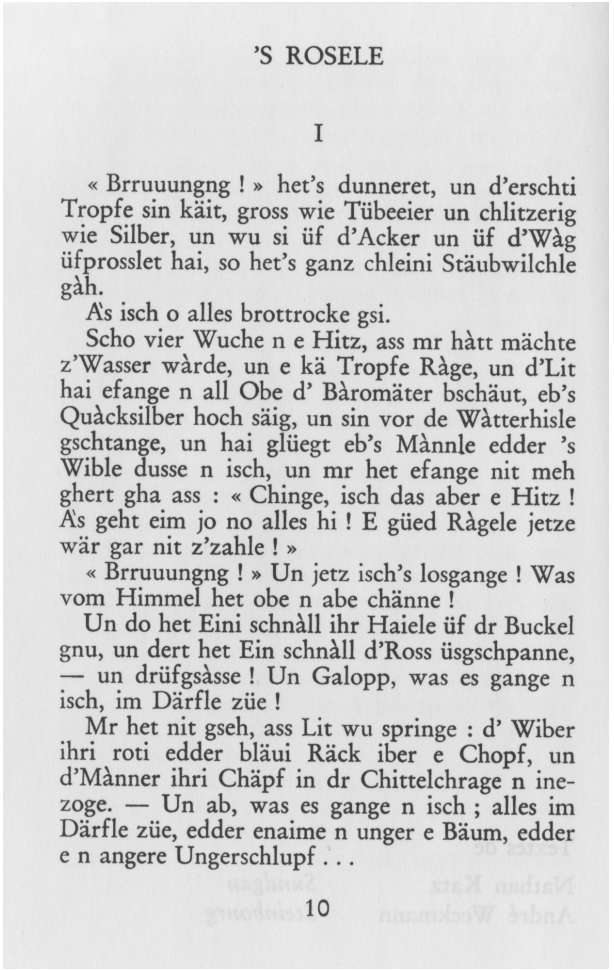
Apprentissage d'un modèle spécifique à la langue :

- Extraction de "traits"
- Utilisation d'un algorithme d'apprentissage

Modèle langue X

Interface de correction en ligne de Jochre

Katz - Mi Sundgäu. Alemannische Gedichte in Sundgauer Mundart, Page 10



’S ROSELE

I

« Brruuungng ! » het’s dunneret, un d’erschti Tropfe sin käit, gross wie Tübeeier un chlitzerig wie Silber, un wu si üf d’Acker un üf d’Wàg üfprosslet hai, so het’s ganz chleini Stäubwilchle gàh.

Às isch o alles brottrocke gsi.

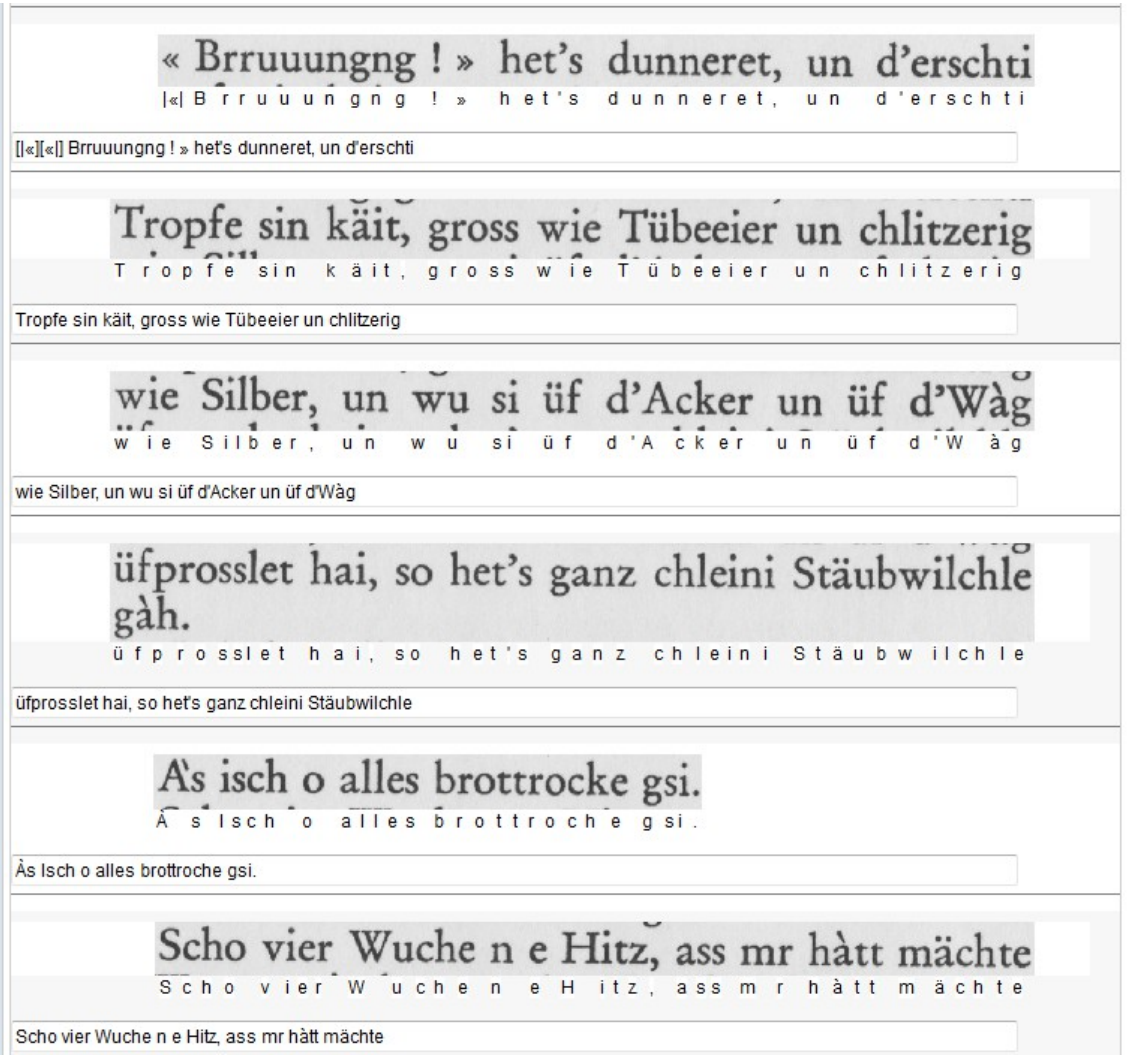
Scho vier Wuche n e Hitz, ass mr hàtt mächte z’Wasser wàrde, un e kä Tropfe Ràge, un d’Lit hai efange n all Obe d’ Bàromäter bschäut, eb’s Quàcksilber hoch säig, un sin vor de Wàtterhisle gschtange, un hai glüegt eb’s Mànnle edder ’s Wible dusse n isch, un mr het efange nit meh ghert gha ass : « Chinge, isch das aber e Hitz ! Às geht eim jo no alles hi ! E güed Ràgele jetze wär gar nit z’zahle ! »

« Brruuungng ! » Un jetz isch’s losgange ! Was vom Himmel het obe n abe chänne !

Un do het Eini schnàll ihr Haiele üf dr Buckel gnu, un dert het Ein schnàll d’Ross üsgschpanne, — un drüfgsàsse ! Un Galopp, was es gange n isch, im Därfle züe !

Mr het nit gseh, ass Lit wu springe : d’ Wiber ihri roti edder bläui Räck iber e Chopf, un d’Mànner ihri Chäpf in dr Chittelchrage n inezoge. — Un ab, was es gange n isch ; alles im Därfle züe, edder enaime n unger e Bäum, edder e n angere Ungerschlupf . . .

10



OCR pour l'alsacien

Outils (open source) :	**Modèles existants**
• **Tesseract :** open source depuis 2005 et développé par Google	• français • allemand • … (plus de 100 langues / typographies)
• **Jochre :** développé par Assaf Urieli	• yiddish et occitan

Dépasser les a priori...

- *"Il n'existe en effet qu'une seule définition scientifiquement correcte de la langue régionale en Alsace, ce sont les dialectes alsaciens dont l'**expression écrite est l'allemand**".*

Recteur Pierre Deyon, 1985

Dépasser les a priori...

- *"Il n'existe en effet qu'une seule définition scientifiquement correcte de la langue régionale en Alsace, ce sont les dialectes alsaciens dont l'**expression écrite est l'allemand**".*

  Recteur Pierre Deyon, 1985

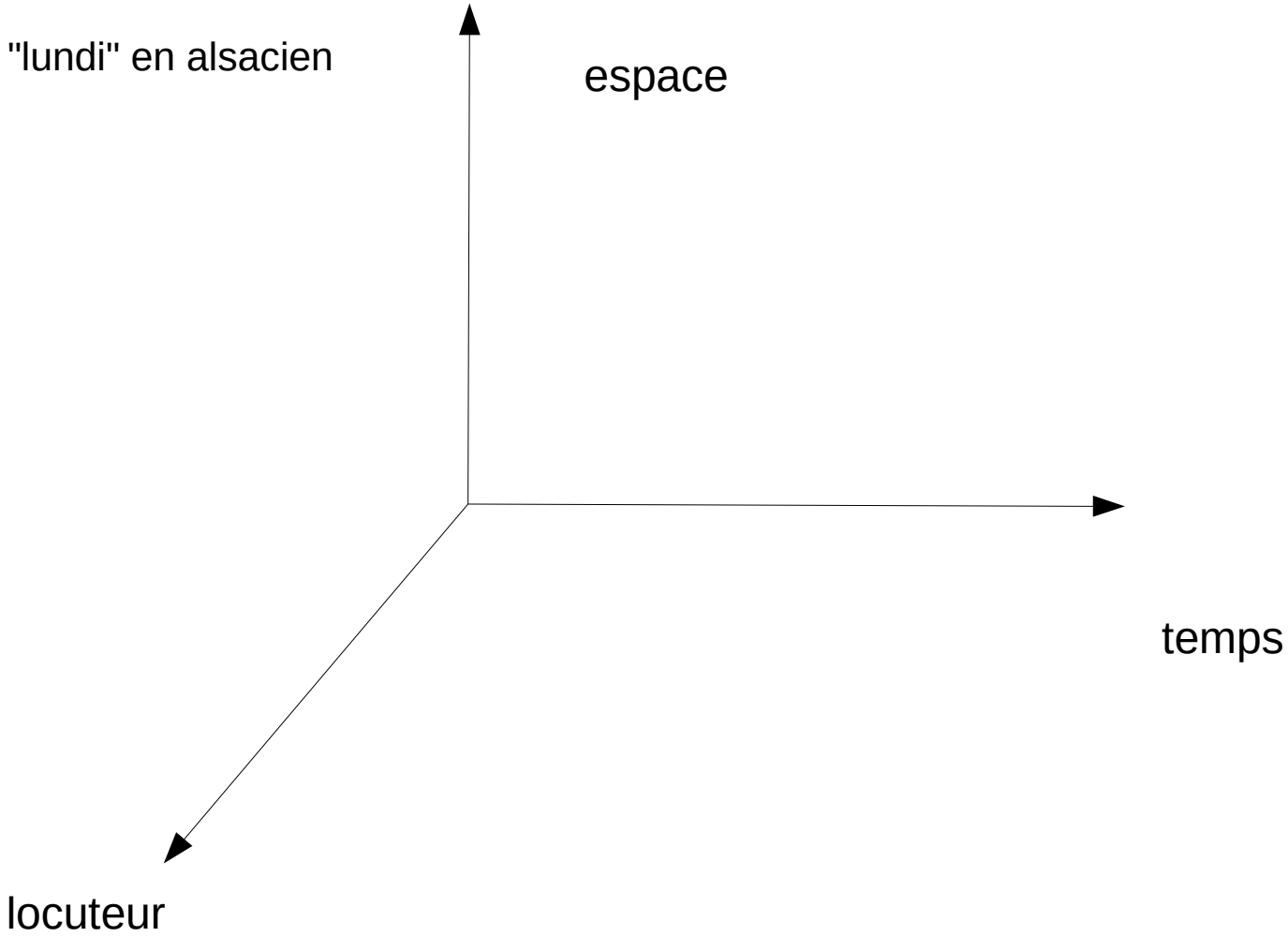
- Résultats pour Tesseract :

Modèle	Pourcentage d'erreurs (caractères)	Exactitude des mots (après normalisation)
allemand	2,33%	**91,25%**
français	2,42%	**91,03%**
français + allemand	1,68%	**94,27%**

- Pourquoi le modèle français fonctionne-t-il aussi bien ?

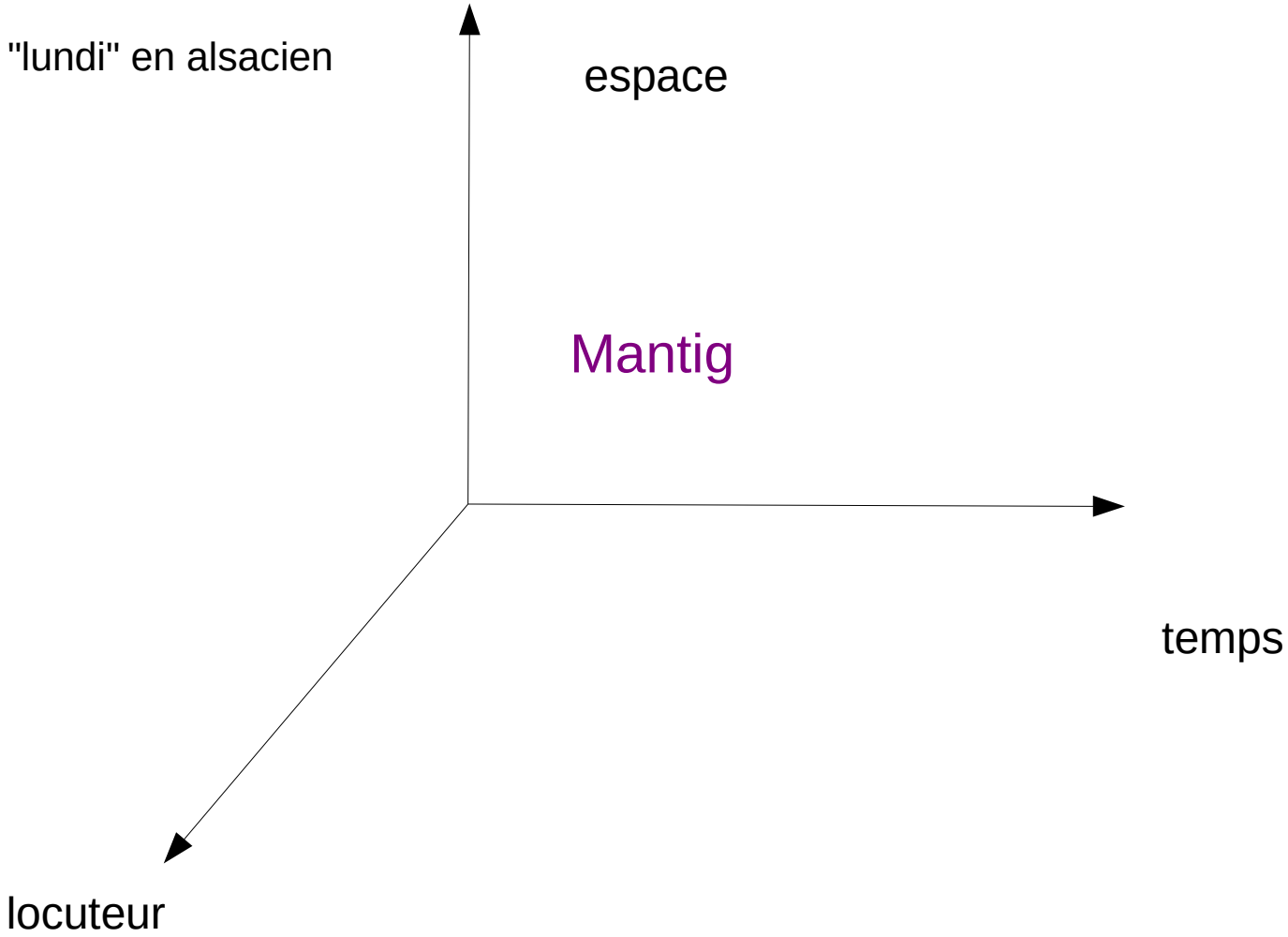
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



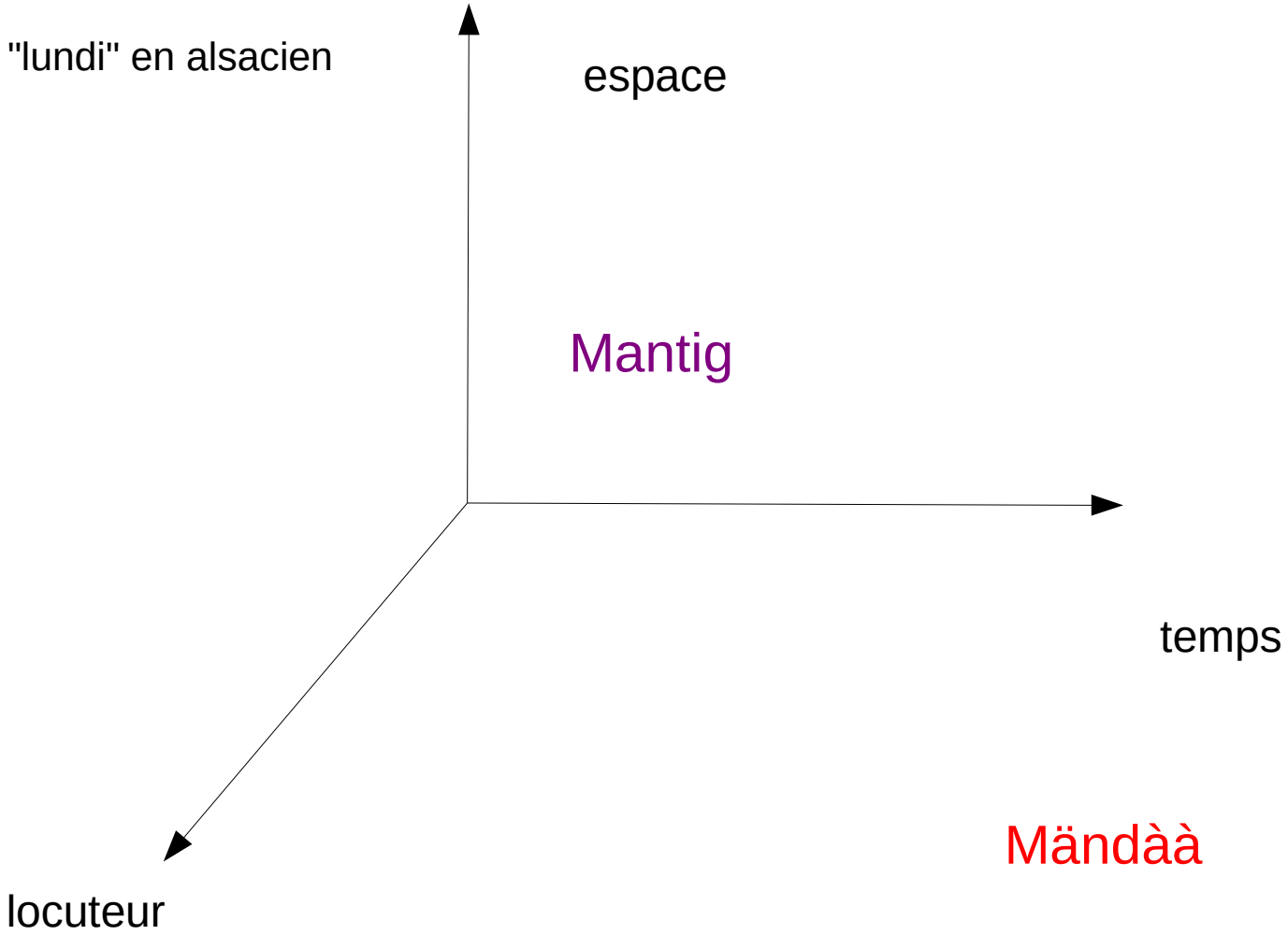
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



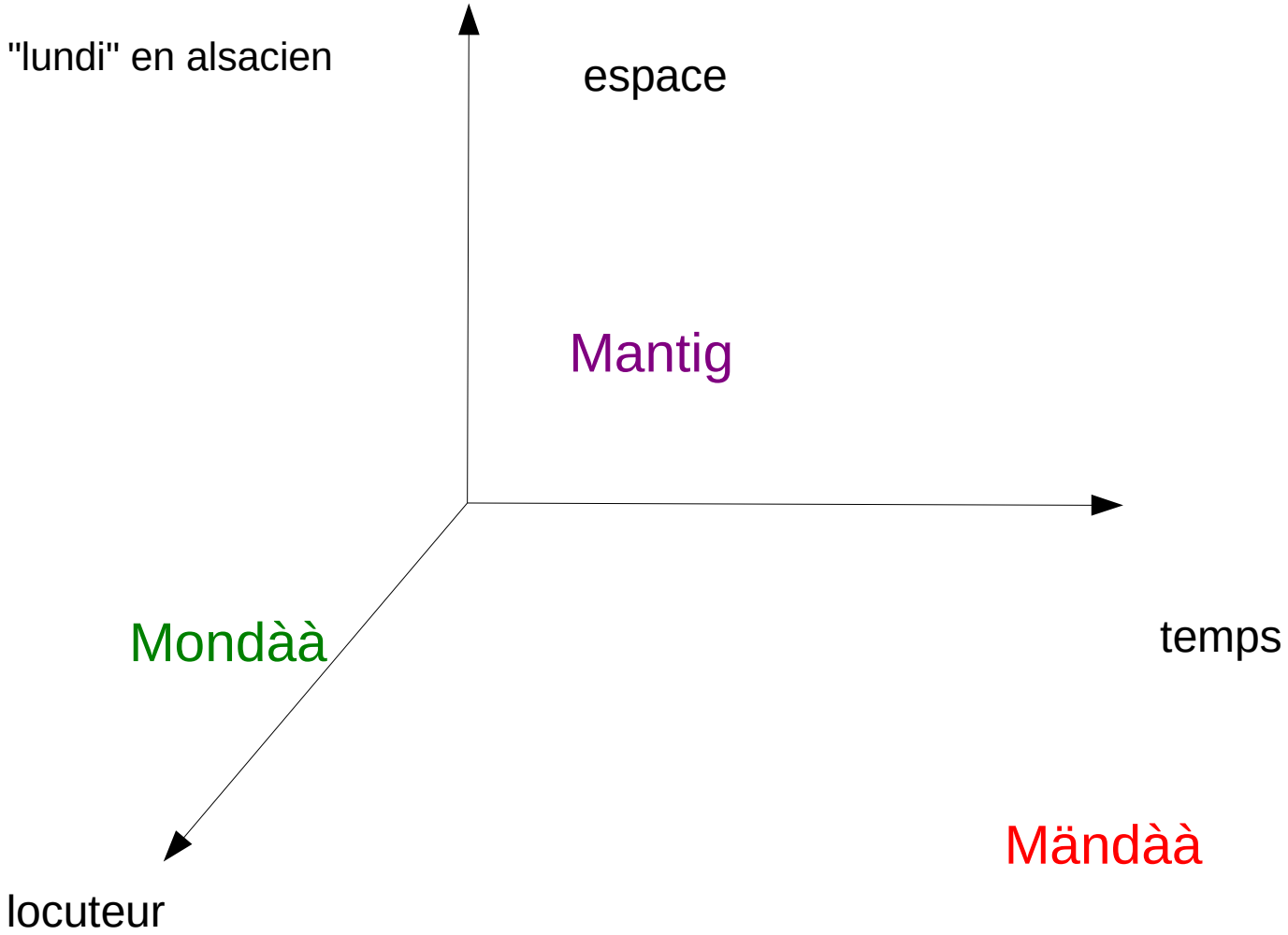
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



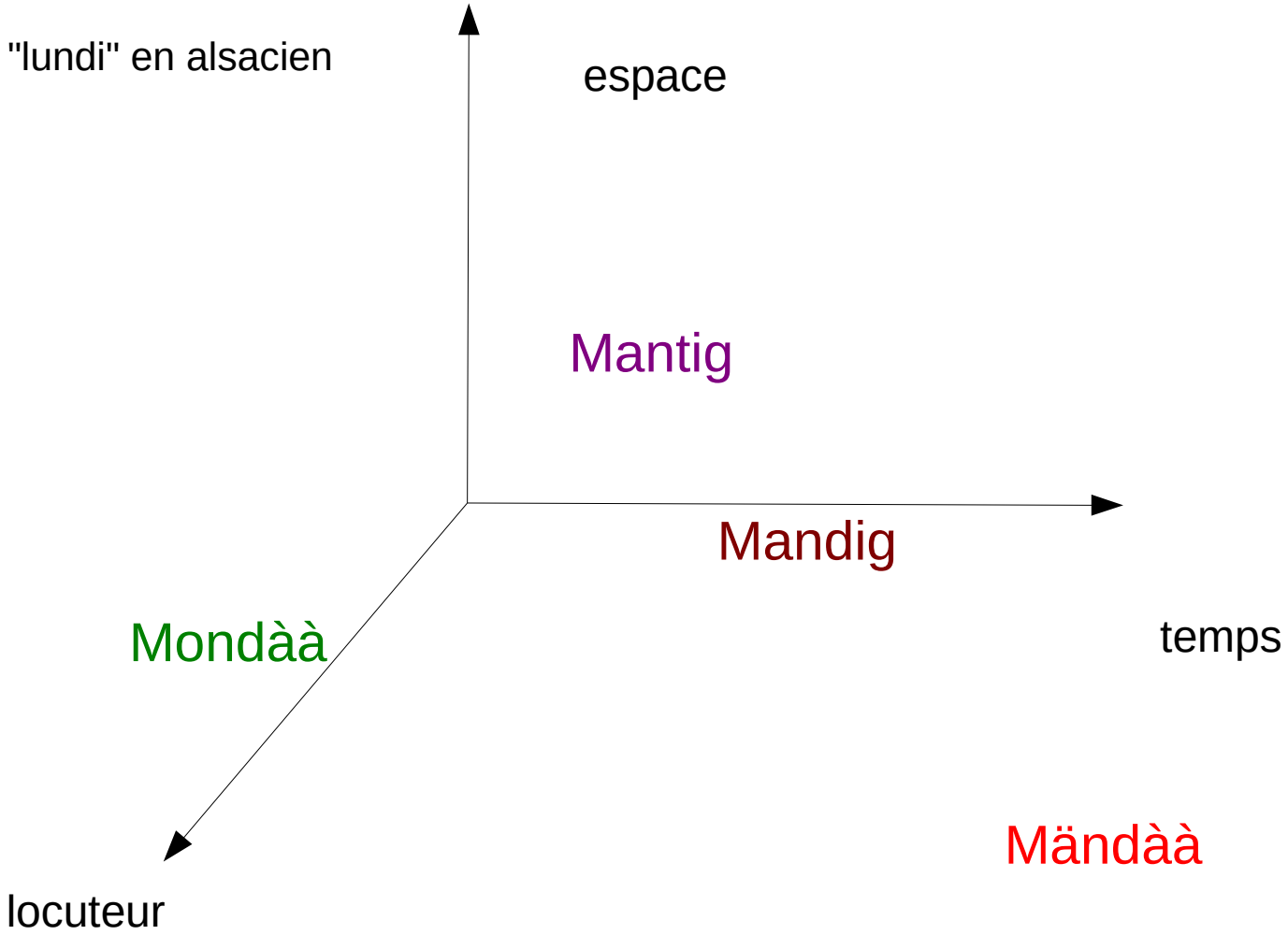
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



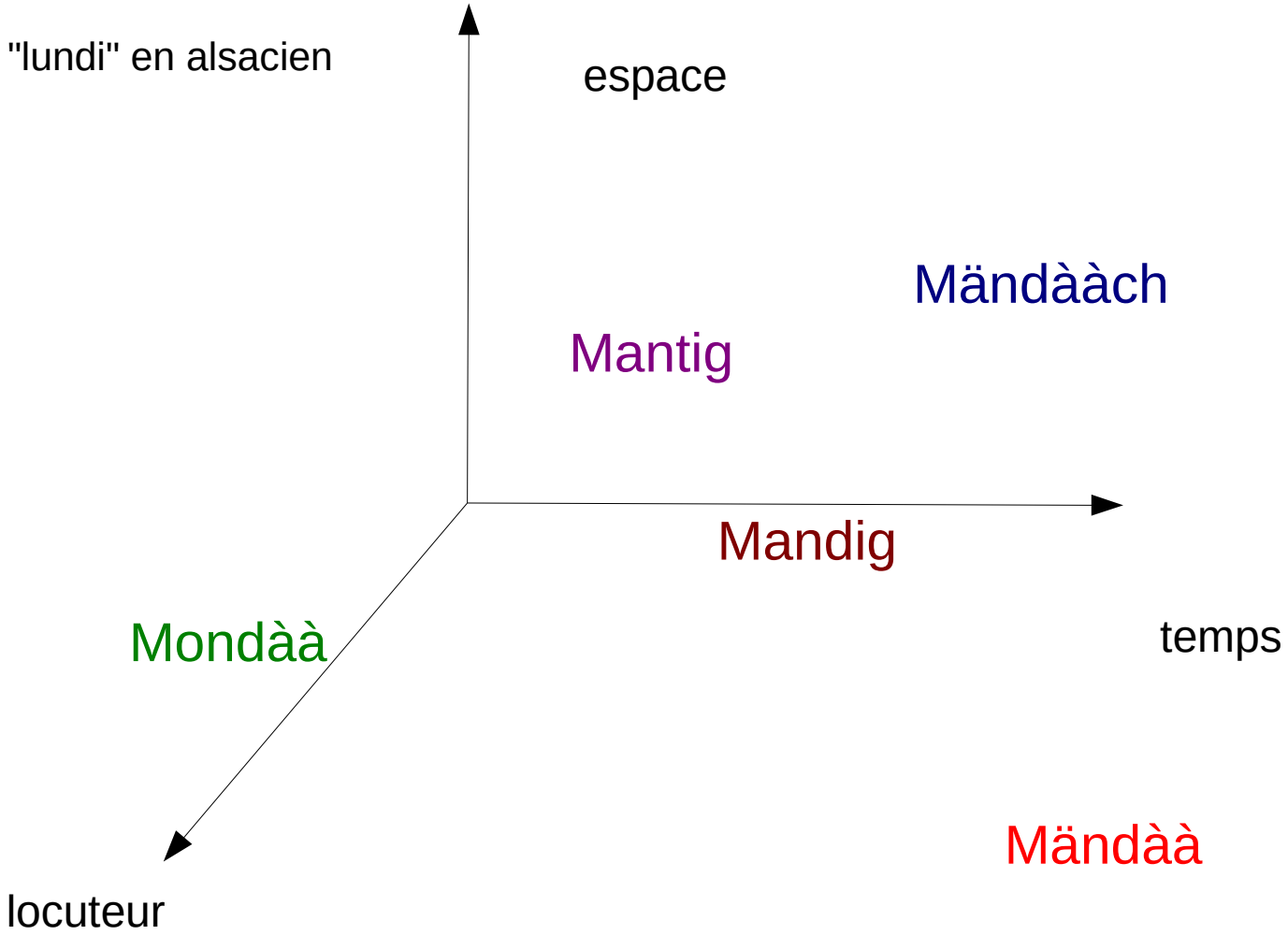
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



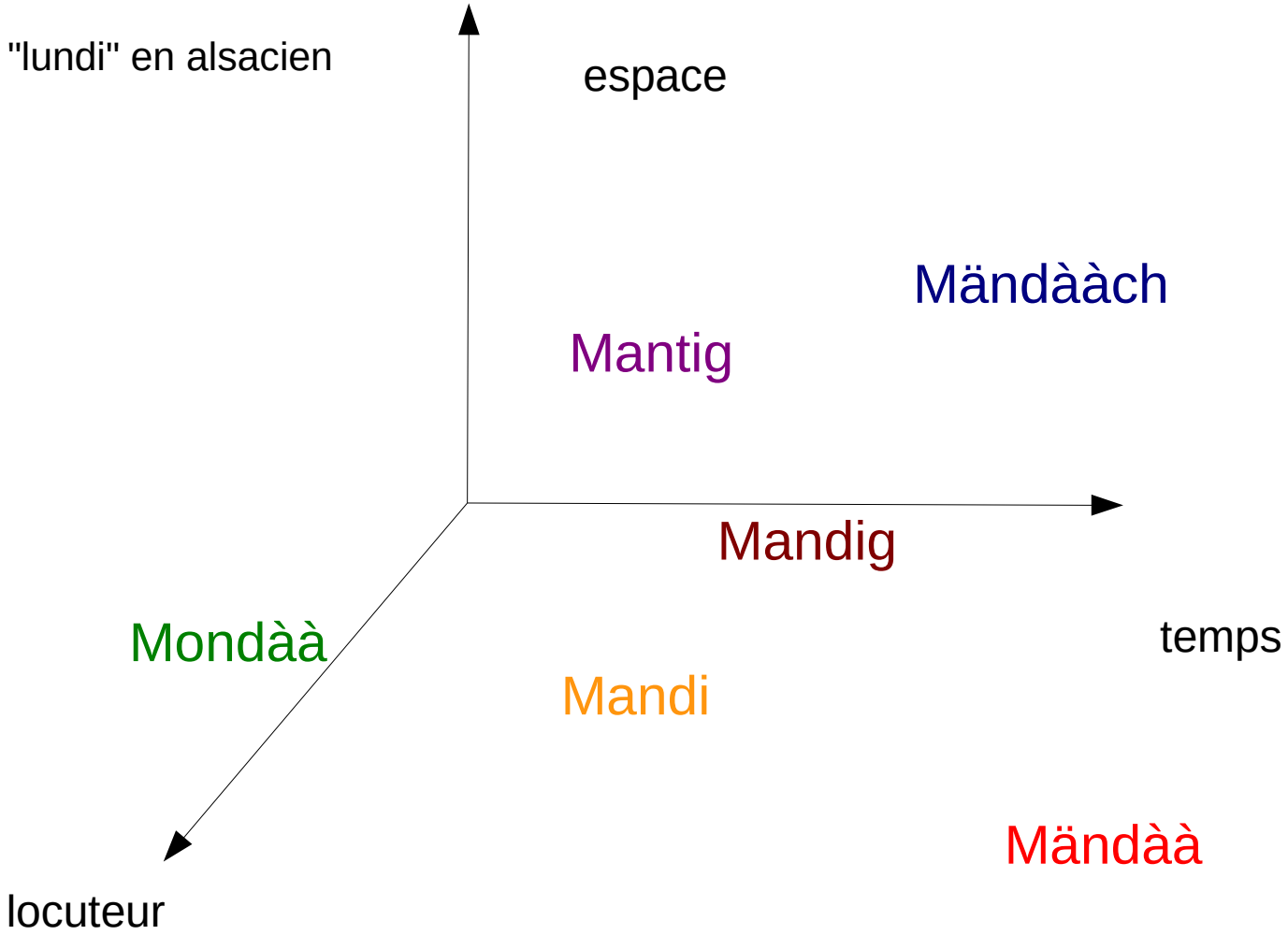
Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



Un défi majeur : la variation graphique

Exemple : "lundi" en alsacien



L'alsacien a du caractère

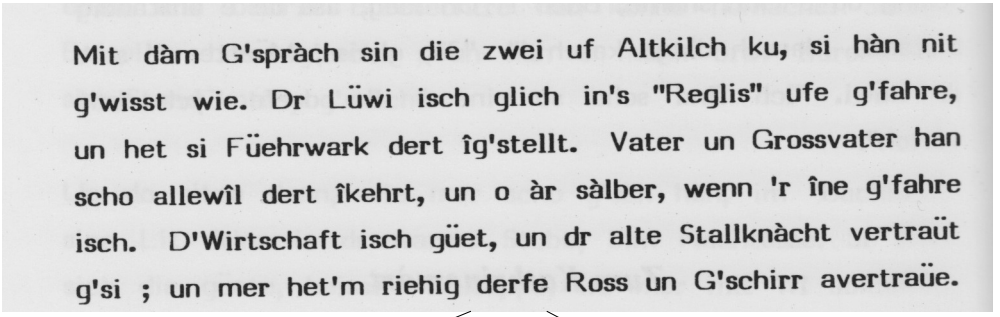
Caractères spécifiquement allemands	Total : 563
ä	419
Ä	5
ö	83
ß	56

Les caractères allemands couvrent **98,3 %** du corpus...

… les caractères français : **99 %**

Caractères spécifiquement français	Total : 960
à	651
À	22
â	75
ç	4
Ç	1
é	154
É	1
è	21
ê	1
î	23
ï	4
ô	2
ù	1

Exemple


Mit dàm G'spràch sin die zwei uf Altkilch ku, si hàn nit g'wisst wie. Dr Lüwi isch glich in's "Reglis" ufe g'fahre, un het si Füehrwark dert îg'stellt. Vater un Grossvater han scho allewîl dert îkehrt, un o àr sàlber, wenn 'r îne g'fahre isch. D'Wirtschaft isch güet, un dr alte Stallknàcht vertraüt g'si ; un mer het'm riehig derfe Ross un G'schirr avertraüe.

'S Bachludis Mariannle, Charles Zumstein, 1986

Modèle Tesseract

allemand

Mit däm G'spräch sin die zwei uf Altkilch ku, si hän nit g'wisst wie. Dr Lüwi isch glich in's "Reglis" ufe g'fahre, un het si Füehrwark dert ig'stellt. Vater un Grossvater han scho allew'il dert 'ikehrt, un o är sälber, wenn 'r ine g'fahre isch. D'Wirtschaft isch güet, un dr alte Stallknächt vertraüt g'si ; un mer het'm riehig derfe Ross un G'schirr avertraüe.

français

Mit dàm G'spràch sin die zwei uf Altkilch ku, si hàn nit g'wisst wie. Dr Lüwi isch glich in's "Reglis" ufe g'fahre, un het si Füehrwark dert îg'stellt. Vater un Grossvater han scho allewîl dert îkehrt, un 0 àr sàlber, wenn 'r îne g'fahre isch. D'Wirtschaft isch güet, un dr alte Stallknàcht vertraüt g'si ; un mer het'm riehig derfe Ross un G'schirr avertraüe.

Perspectives

- Poursuite du travail sur l'OCR pour améliorer les résultats
  - Entraînement de modèles pour l'alsacien
  - Combinaison avec les modèles existants

- Prochaine application visée : annotation morpho-syntaxique
  - Définition de jeux d'étiquettes
  - Annotation de corpus pour l'évaluation
  - Constitution de lexiques morphosyntaxiques
  - Transfert technologique d'outils existants

Merci vielmols !