A. Dawson, J-M. Eloy, C. Rey

LESCLaP (CERCLL)

Université de Picardie Jules Verne

a.dawson@free.fr

jean-michel.eloy@u-picardie.fr

christophe.rey@u-picardie.fr

Colloque *de l'AFLS*

8-10 septembre 2011

Nancy

Vue perspective sur le français à partir d'une base de données textuelles en domaine d'oïl

Quelques données-clés pour présenter Picartext

Le modèle de la base FRANTEXT →

- Une base textuelle comprenant actuellement environ 10 millions de mots
- Une base textuelle littéraire composée d'éléments divers (dictionnaires, contes, recueils de poésies, romans, chansons, etc.)
- Une base panchronique
  - ✓ "moderne" depuis le XVII^{e} siècle
  - ✓ "médiévale"

Les objectifs de Picartext

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

Aspects linguistiques

- Explorations de la langue très diverses : lexique, grammaire, phraséologie, styles, etc. *=>Extraction de sous-produits dictionnairiques divers (par domaines, par constructions, par lieux, dates, genres, etc.)*

- Étude des évolutions de la langue, connectées aux lieux et aux biographies des auteurs

=>faire émerger le commun - unitaire (koïnè) – élaborer la standardisation, élaboration de la norme graphique

- Outils informatiques «robustes» de traitement de la variation

Aspects culturels

- Soutien à la création littéraire, aux jeux, etc.

UNIVERSITÉ de Picardie Jules Verne

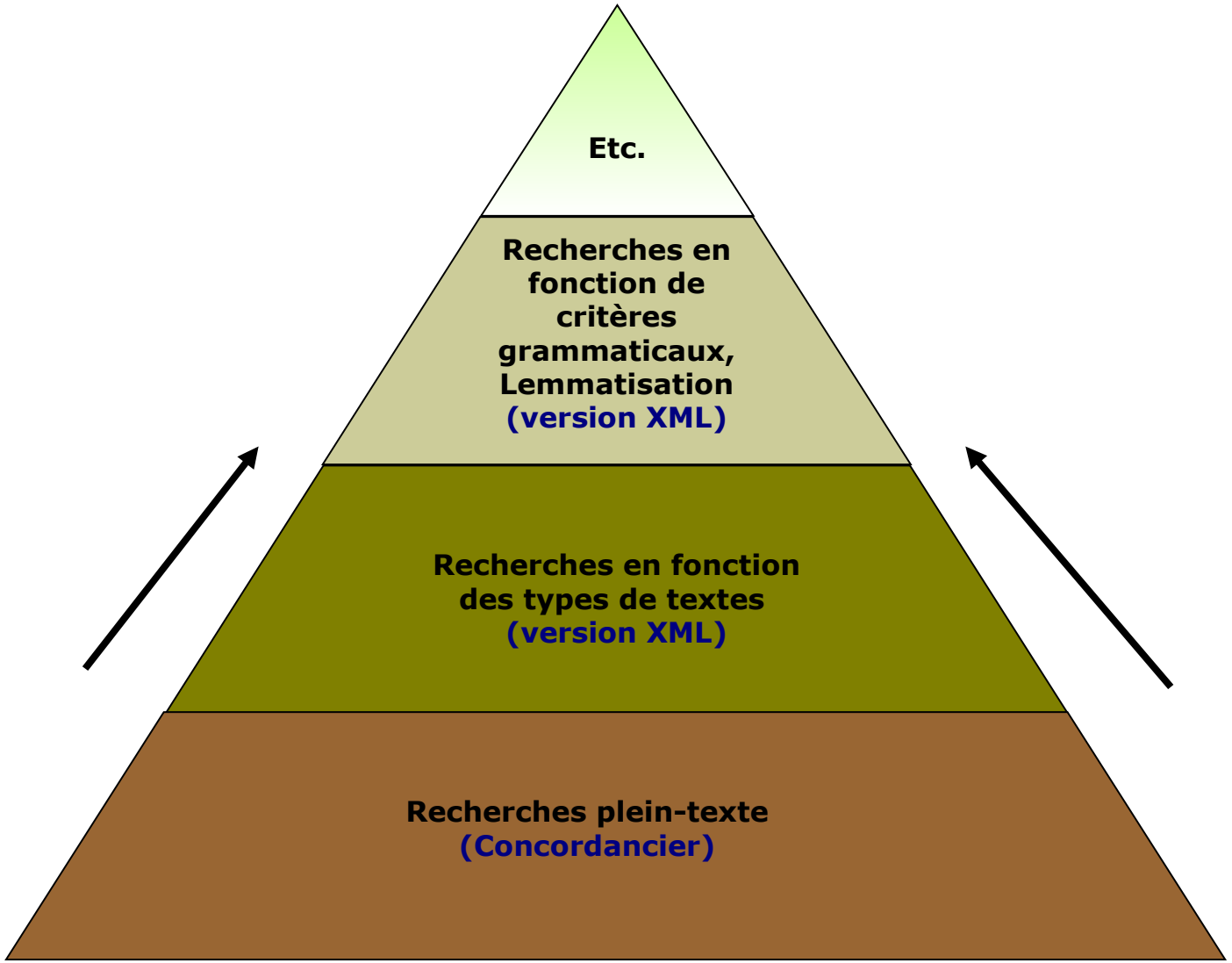
Exploitation de la base Picartext

Proposer un éventail de fonctionnalités

En préparation

Bientôt disponible

Disponible



Un outil accessible en ligne

- **http://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/index.php**

UNIVERSITÉ de Picardie Jules Verne

LESCLaP
Laboratoires d'Etudes Sociolinguistiques
sur les Contacts de Langues et la Politique Linguistique

CEP CENTRE D'ETUDES PICARDES

PICARDIE LA RÉGION

Bienvenue sur le site du Projet PICARTEXT

Accès public aux données PICARTEXT

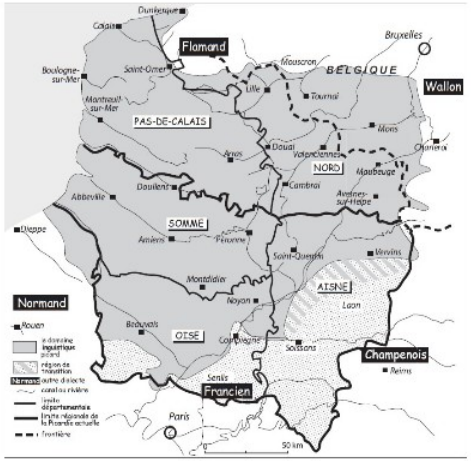
Présentation du Projet PICARTEXT

La base Picartext, développée par le Laboratoire d'Études Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique (CERCLL-LESCLaP – EA 4283) de l'Université de Picardie Jules Verne avec le soutien financier de la Région de Picardie, est constituée de textes écrits partiellement ou totalement en picard, issus de l'ensemble du domaine linguistique picard, et composés depuis le XVIII[e] siècle jusqu'à nos jours.
L'objectif est d'offrir à la communauté des chercheurs, ainsi qu'à un public averti, une ressource linguistique à partir de laquelle il sera possible d'envisager toutes sortes d'exploitations :

- exploration de la langue (lexique, morpho-syntaxe, phraséologie...)
- étude des évolutions diachroniques
- étude de la variation et de la cohésion dialectales et du processus de koinèisation

À partir de Picartext pourront être créés des outils à disposition des usagers de la langue :

- dictionnaires appuyés sur les usages
- outils pédagogiques
- standardisation graphique



Le domaine linguistique picard

Des données vers la théorie (1)

langues collatérales « des variétés proches – objectivement et subjectivement –, aux plans linguistique, sociolinguistique et historique ou glottopolitique, les variétés tendanciellement en contraste étant historiquement liées par les modalités de leur développement ».

"**variétés**" peut désigner ce que les sociétés nomment "*langues, dialectes, patois, modalités, etc*." ;

"**développement**" fait allusion à la construction et à l'élaboration des langues (cf. Kloss : *Ausbau*, mais aussi *grammaticalisation*, *normes*);

"**historiquement**" réfère à l'histoire sociale, ou encore à la dynamique sociolinguistique ; "dynamique" rappelle que ces variétés sont toujours mouvantes ;

"**lié**" peut renvoyer à l'attraction ou à la répulsion, le plus souvent coexistantes dans les pratiques ;

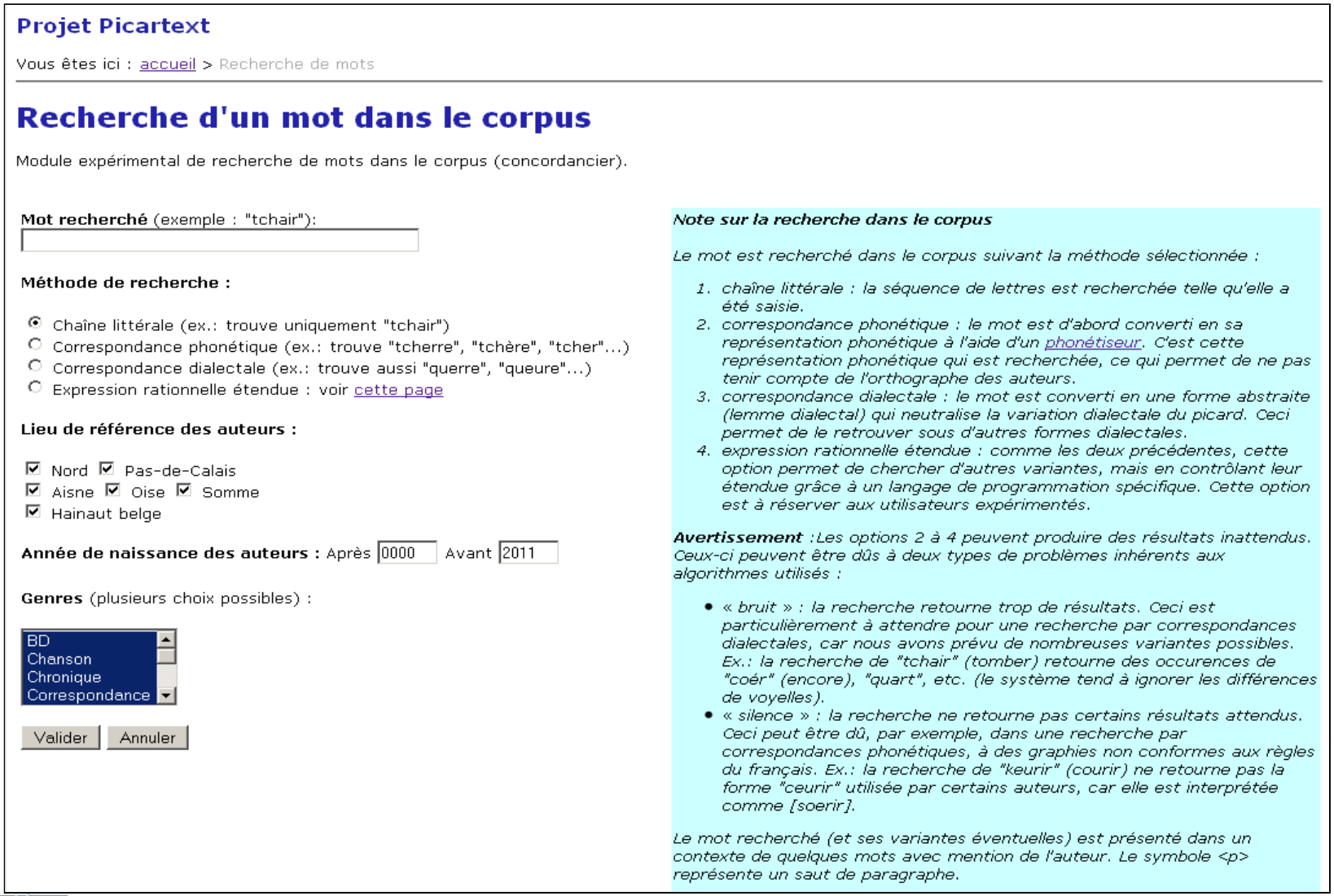
"**complexe**" doit s'entendre au sens de la "pensée complexe" (E. Morin 1990) . En très bref les processus complexes sont non mécanistes, en particulier par l'intervention des représentations ou épilinguistique

De la théorie vers les données

- dans la constitution du concordancier
-Théorie des Correspondances
- généralisation : continuités / ruptures

Concordancier dialectal (définition de la requête)

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique



Projet Picartext

Vous êtes ici : accueil > Recherche de mots

Recherche d'un mot dans le corpus

Module expérimental de recherche de mots dans le corpus (concordancier).

Mot recherché (exemple : "tchair"):

Méthode de recherche :

- (•) Chaîne littérale (ex.: trouve uniquement "tchair")
- ( ) Correspondance phonétique (ex.: trouve "tcherre", "tchère", "tcher"...)
- ( ) Correspondance dialectale (ex.: trouve aussi "querre", "queure"...)
- ( ) Expression rationnelle étendue : voir cette page

Lieu de référence des auteurs :

☑ Nord ☑ Pas-de-Calais
☑ Aisne ☑ Oise ☑ Somme
☑ Hainaut belge

Année de naissance des auteurs : Après 0000 Avant 2011

Genres (plusieurs choix possibles) :

BD
Chanson
Chronique
Correspondance

Valider Annuler

Note sur la recherche dans le corpus

Le mot est recherché dans le corpus suivant la méthode sélectionnée :

1. *chaîne littérale : la séquence de lettres est recherchée telle qu'elle a été saisie.*
2. *correspondance phonétique : le mot est d'abord converti en sa représentation phonétique à l'aide d'un phonétiseur. C'est cette représentation phonétique qui est recherchée, ce qui permet de ne pas tenir compte de l'orthographe des auteurs.*
3. *correspondance dialectale : le mot est converti en une forme abstraite (lemme dialectal) qui neutralise la variation dialectale du picard. Ceci permet de le retrouver sous d'autres formes dialectales.*
4. *expression rationnelle étendue : comme les deux précédentes, cette option permet de chercher d'autres variantes, mais en contrôlant leur étendue grâce à un langage de programmation spécifique. Cette option est à réserver aux utilisateurs expérimentés.*

***Avertissement** :Les options 2 à 4 peuvent produire des résultats inattendus. Ceux-ci peuvent être dûs à deux types de problèmes inhérents aux algorithmes utilisés :*

- *« bruit » : la recherche retourne trop de résultats. Ceci est particulièrement à attendre pour une recherche par correspondances dialectales, car nous avons prévu de nombreuses variantes possibles. Ex.: la recherche de "tchair" (tomber) retourne des occurences de "coér" (encore), "quart", etc. (le système tend à ignorer les différences de voyelles).*
- *« silence » : la recherche ne retourne pas certains résultats attendus. Ceci peut être dû, par exemple, dans une recherche par correspondances phonétiques, à des graphies non conformes aux règles du français. Ex.: la recherche de "keurir" (courir) ne retourne pas la forme "ceurir" utilisée par certains auteurs, car elle est interprétée comme [soerir].*

Le mot recherché (et ses variantes éventuelles) est présenté dans un contexte de quelques mots avec mention de l'auteur. Le symbole <p> représente un saut de paragraphe.

UNIVERSITÉ de Picardie Jules Verne

Concordancier dialectal (page de résultats)

Recherche d'un mot dans le corpus

Module expérimental de recherche de mots dans le corpus (concordancier).

Mot recherché : sodar

Méthode de recherche : Correspondance dialectale
Lieu de référence : non pris en compte
Année de naissance : entre 0000 et 2011
Genre(s) : non pris en compte

iez leu éporons,<p>27.Et y quitte Ieu garnigeons,<p>28.Le	*Saudar*	suit sen Capitaine,<p>29.Sans rien savoir du qu'un le me	(Jacques DECOTTIGNIES)
t en méme-tems pour no Ville,<p>11.Car cha fégeoit un bon	*Saudar*	,<p>12.Regrettez du tier & du quart,<p>13.Quoi faire Dieu	(Jacques DECOTTIGNIES)
ausy vive que n'est de le poudre,<p>42. Il a rasanné ses	*Saudards*	10,<p>43. Pour vire qu'il aroit mengé le lard11,<p>44. To	(Jacques DECOTTIGNIES)
het ichy qui faut tenir tiete, [p.3]<p>53. Nos ROY, ses	*Saudards*	, tout s'aprête, <p>54. Et un plache tout nos Quennons, <	(Jacques DECOTTIGNIES)
it le dessus,<p>110.Un a crié VIVE LE ROY,<p>111.Tous nos	*Saudards*	étoite en joye30 :<p>112.Un ne peut point vous en dire l	(Jacques DECOTTIGNIES)
p>147.Le ROY, le DOPHIN, les Générales, <p>148. Aveuq leu	*Saudars*	sans delay, <p>149. Ont revenu tout prez de Tournay40, <	(Jacques DECOTTIGNIES)
<p>180.D'aler tanté un co de se main.<p>181.Vela tout nos	*Saudars*	en quemain,<p>182.Aveuq leu fusique amorsé,<p>183.D'un g	(Jacques DECOTTIGNIES)
aiche dormire leu éventelles,<p>22.Que les Chentinelles &	*Saudars*	,<p>23.Ont des Capotes sur les Ramparts,<p>24.Il a surven	(Jacques DECOTTIGNIES)
Ramparts,<p>24.Il a survenuë unne allerte,<p>25. Tous nos	*Saudars*	mette leu guettes,<p>26.Nos Cavaliez leu éporons,<p>27.E	(Jacques DECOTTIGNIES)
03. Enportoint leu pain et leu crache40,<p>104. Et sy nos	*Saudars*	sont ensain,<p>105. Ch'est qui leu on moutrez le quemin.	(Jacques DECOTTIGNIES)
7. Velà le Ville bétot prête à rendre.<p>178.Grament de	*Saudars*	un envoye<p>179.Et le Regiment du fieu du Roy69,<p>180.P	(Jacques DECOTTIGNIES)
l'a bétot vû su le rempart70 <p>187.Sy vous arite vû nos	*Saudars*	, <p>188.Quand y ont aperchu l'ensaine <p>189.L'ont moutr	(Jacques DECOTTIGNIES)
toit le mignon de Vienne,<p>258.Tout cha fegeoit des bons	*Saudars*	<p>259.Che-ty-chy va encore à part,<p>260.Ch'est grament	(Jacques DECOTTIGNIES)
re chel Ville<p>261.Wardé d'environ quinze mille<p>262.De	*Saudars*	sans les cras pingeons92 <p>263.Qu`un r'envoye dessus le	(Jacques DECOTTIGNIES)
chellale, <p>263.Conty envoye le princhipale<p>264.De ses	*Saudars*	en pau pus long,<p>265.Ne wardant que six Battaillons,<p	(Jacques DECOTTIGNIES)
p>29.Venoit de faire trois chens lieuës,<p>30.Aveuque ses	*saudars*	à moustache,<p>31.Pour vire nos Franchois fache à fache:	(Jacques DECOTTIGNIES)
pau avant,<p>127.Un envoy un Prinche du sang29,<p>128.Des	*saudars*	y n'y avoit qu'a prendre,<p>129.Deden les dessain[ ]d'en	(Jacques DECOTTIGNIES)
[p.7]<p>131.Tout le pu bielle du Namurois.<p>132.Nos	*saudars*	passe ses rivieres30,<p>133.Pour l'entourez devant derri	(Jacques DECOTTIGNIES)
coûteummé<p>151.D'être aveuque les bras croijé,<p>152.Nos	*saudars*	étointe deja mate<p>153.De ne pu avoir de bled à batte,<	(Jacques DECOTTIGNIES)

Concordancier dialectal : traitement de la variation

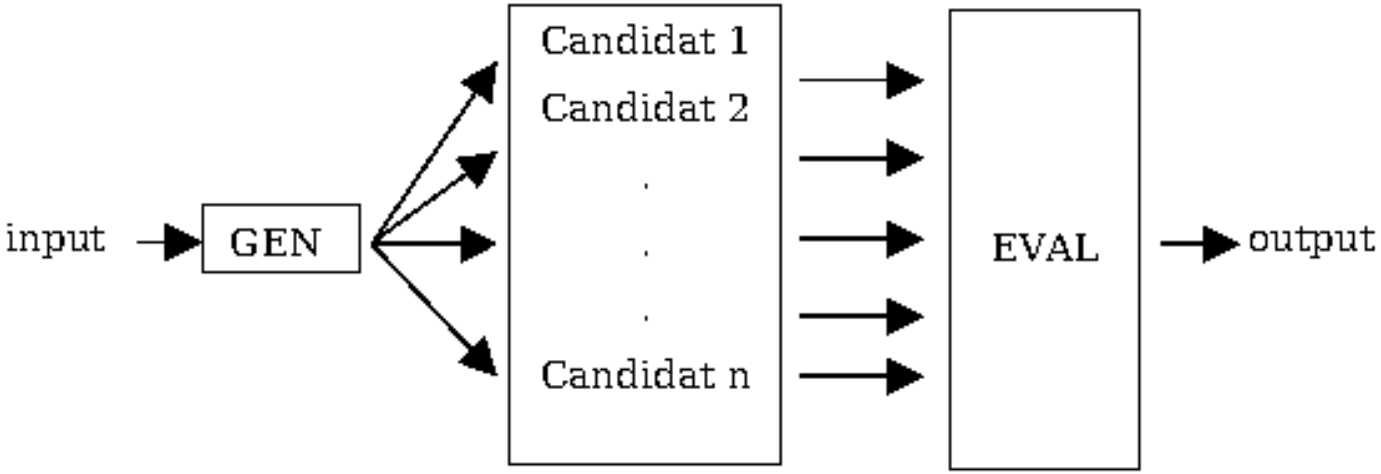
Variation graphique

- Constat : les règles grapho-phoniques du français sont généralement valables pour le picard *(quelques exceptions)*
- Utilisation d'un phonétiseur dérivé de TTS-French développé dans le cadre du projet MBROLA (Faculté Polytechnique de Mons)
- Indexation du corpus à l'aide de transcriptions phonétiques (ou plutôt phonologiques) générées automatiquement
- Une recherche de forme retourne toutes ses occurrences, quelles que soient les graphies adoptées par les auteurs

Variation phonologique

- Indexation à l'aide de lemmes dialectaux subsumant la variation phonologique
- Basée sur la Théorie des Correspondances Dialectales (Dawson 2006)

Schéma d'une grammaire OT (1)



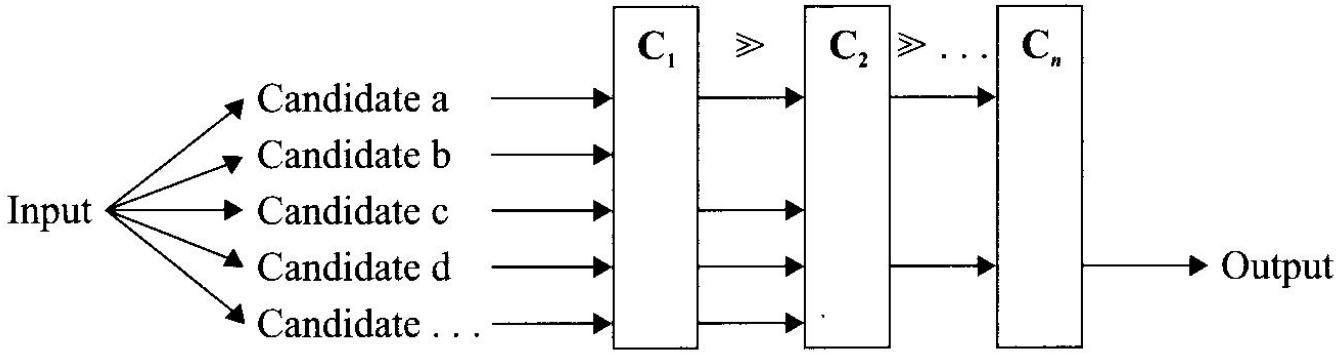
OT est une grammaire générative.

Le module génératif ou générateur (GEN) engendre une *infinité de candidats* à partir de l'input

La fonction d'évaluation ou évaluateur (EVAL) sélectionne le *candidat optimal* = output

EVAL fait appel à un *ensemble ordonné de contraintes* = CON

Schéma d'une grammaire OT (2)



EVAL met en jeu un ensemble ordonné de contraintes.
Les contraintes sont essentiellement de deux types :
- contraintes de marque
- contraintes de fidélité (correspondance) :
« Etant donné deux chaînes S1 et S2, la **correspondance** est une relation R des éléments de S1 sur les éléments de S2. Deux éléments α ∈ S1 et β ∈ S2 sont désignés comme **correspondants** lorsque α R β. »

Contraintes de correspondance

MAX-IO : Tout segment de S1 a un correspondant dans S2 : S1 se projette « maximalement » sur S2 (contrainte anti-effacement)

DEP-IO : Tout segment de S2 a un correspondant dans S1 : S2 « dépend » de S1 (contrainte anti-épenthèse)

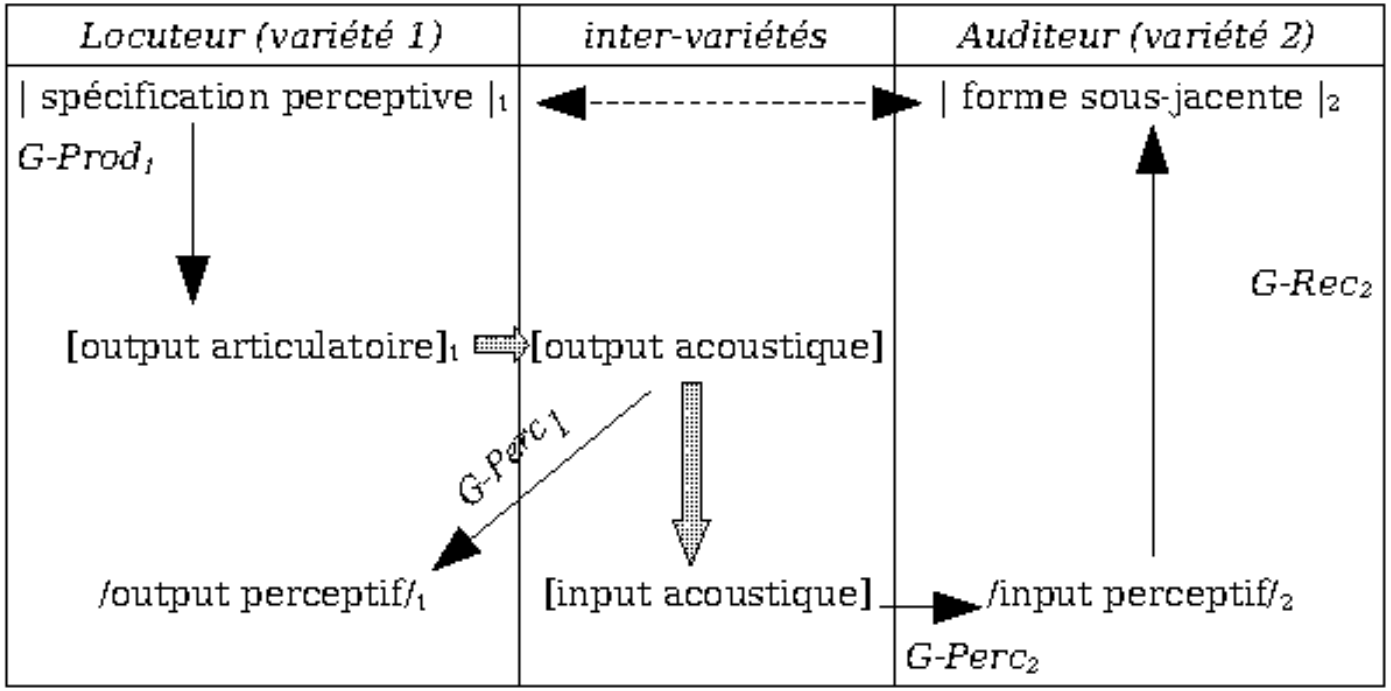
Ident-IO(F) : Les segments correspondants doivent être identiques vis-à-vis du trait F

Correspondances Dialectales (1)

- Implémentation de la « force d'intercourse » (Saussure) : contraintes de correspondance inter-variétés : Max-V_1V_2, Dep-V_1V_2, Ident-V_1V_2
- Elles visent la forme sous-jacente de l'énonciataire (V_2) et contrôlent la conformité de l'output articulatoire avec la représentation que s'en fait le locuteur

Correspondances Dialectales (2)

Grammaire d'interlocution en situation inter-variétés (d'après Boersma 1998, 1999)



Correspondances dialectales(3)

Exemple de contrainte de correspondance dialectale :
*Distordre-V_1V_2([trans.F2/F3] : 1)

Tableau des distorsions :

	k [+arrière] 2	k [–arrière] 3	kʲ 4	ʧ 4
k [+arrière] 2	0	1	2	2
k [–arrière] 3	1	0	1	1
kʲ 4	2	1	0	0
ʧ 4	2	1	0	0

Correspondances dialectales (4)

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

|ko| « chaud » V_2 :	ko		*o	*Distordre-V_1V_2([trans. F2/F3] : 1)	Aligne-G(cor', σ)	*StrComp	Ident-IO(dor)
[ko]	*!	(d=0)			
[kʲo]	*!	* (d=2)		*	
[ʧo]	*!	* (d=2)			*
. [kø]		(d=1)	*		*
[kʲø]		*! (d=2)		*	*
[ʧø]		*! (d=2)			**

|kø| « queue » V_2 :	kø		*o	*Distordre-V_1V_2([trans. F2/F3] : 1)	Aligne-G(cor', σ)	*StrComp	Ident-IO(dor)
[ko]	*!	(d=1)			*
[kʲo]	*!	(d=1)		*	*
[ʧo]	*!	(d=1)			**
[kø]		(d=0)	*!		
[kʲø]		(d=1)		*!	
. [ʧø]		(d=1)			*

Correspondances dialectales (5)

- Elles rendent compte de phénomènes d'identification inter-variétés (grammaire de compréhension),
- et des facteurs de cohésion (grammaire de production)
- Ancrées dans les données acoustiques et perceptives → limites de la variation
- Exemple : palatalisation secondaire des vélaires en picard, mais pas (non-)palatalisation à l'échelle du domaine d'oïl

Concordancier dialectal : correspondances dialectales

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

- Les variantes dialectales des variétés en situation d'intercommunication sont mises en relation au niveau segmental. Exemple :

K → tS

e → E

r → R

- On en déduit des classes de segments équivalents (exemple : {k, tS}), qui permettent de construire les lemmes dialectaux abstraits
- Exemple de lemme dialectal : //{k,tS}{e,E,a,o…}{r, R}// simplifié en //tSOR//
- Le sous-corpus est indexé à l'aide des lemmes dialectaux de façon identique aux représentations phonétiques, à l'aide du même phonétiseur assorti d'une base de règles spécifique

UNIVERSITÉ de Picardie Jules Verne

Synthèse : théorie → données

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

- enjeu : intégrer dans l'outil la connaissance empirique sur la variation dialectale
- rôle de la théorie : informer cette connaissance pour son implémentation informatique
- la théorie est elle-même ancrée dans des données de bas niveau : données → théorie → données
- /ker/ - /tSER/ vs. Choir : rupture née de potentialités différentes d'identification et d'intercompréhension
- étudier au cas par cas ces potentialités, en complémentarité avec l'éclairage sociolinguistique

Des données vers la théorie (2)

ch'soleu **i** luit - le soleil il luit

Julie Auger (1993) : "It is usually agreed that Picard lexical subjects **always** cooccur with a coreferential subject clitic"

en **français parl**é, Colette Jeanjean (1981) : **les 2/3** des sujets à forme lexicale ("le N") apparaissent en couplage avec un clitique (Ex. "**les types ils savaient plus**").

et dans son corpus de français parlé, **92 %** des sujets de verbe sont des clitiques.

Pour Hrkal (1911) ou Debrie (1974) : c'est une **règle catégorique** en picard.

or même dans leurs exemples , cette règle n'est **pas toujours respectée**.

>>> **site de variation ?** (auquel cas le picard = "le français" ?) **Non**.

"In many cases, it is quite clear that **this is attributable to the influence of Standard French**".

Ainsi, **le recours à l'interférence exolingue a permis de " sauver " une règle catégorique**, et d'**éviter de devoir constater la variation et l'absence de frontière** entre picard et français sur ce point

la démarche de corpus doit être menée sur du français, du picard, de l'intermédiaire, du mélange

Des données vers la théorie (3)

chez le dialectologue Debrie (1974), natif de la région, à propos d'un autre fait :

Debrie : la tournure *"**la toupie que les enfants jouent avec**"* est typique du picard, où le relatif "dont" a disparu,

dans une copie d'élève **en <u>français</u>** , il lui donne le statut d'une **interférence** :

"*elle est le type même de la traduction, celle du mot à mot, qui conserve dans la "langue étrangère" (pour un certain nombre de nos élèves le français est plus souvent qu'on ne le croit une langue inconnue) les tournures et les locutions de la langue "maternelle". "* (Debrie 1974:156).

ainsi , **constatant** dans un discours **français** le même fait "**picard**"

il lui faut **affirmer l'interférence**, la " mauvaise traduction " de " langue étrangère ", **pour maintenir l'étanchéité de principe**, la différence de grammaire.

le <u>but</u> (non conscient) de son analyse : **construire une frontière ferme et nette**

c'est **la description constructive**

de tels gestes rejoignent ce qu'on appelle les attitudes "puristes"

Des données vers la théorie (4)

LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

(J. Auger) **extraction du sujet + proposition relative**, on trouve " systématiquement " le **clitique sujet** est " **presque toujours présent** ".

" **Mi qu'j'o dvini pourquo** " (*moi qui ai deviné pourquoi*)

" l'graind-route **qu'alle traverche** ech village" (*la grand-route qui traverse le village*).

***

Guiraud (1965) : " **décumul du relatif** " en **français populaire**,
" **Elle est là qu'elle attend** ", " **C'est moi que je...** ", etc.

J. Auger : oui, cela existe en " colloquial French "
Mais les **faits picards** s'opposent au " **pronom relatif *qui* du français** "
ce qui renvoie au **contraste figé**

***

exagération de la spécificité du picard, sur ce point : déjà chez **Flutre** (1970 et 1955)
"**V'la l'cloque qu'al sonne**" (" *voilà la cloche qui sonne* "),
de "**constructions tout à fait courantes en picard actuel, mais ignorées du français**".

Réponse : le corpus ! + analyse sociolinguistique de ce qu'est le français standard

Des données vers la théorie (5)

pour + N sujet + infinitif

(le chien accidenté) falleut même el porter **pour li picher**
il fallait même le porter [pour lui pisser] pour qu'il pisse

et pi **pour chés gins rintrer** sins rouvrir él grille
et [pour les gens rentrer] pour que les gens rentrent sans rouvrir la grille

i soulièfe és casquette **pour li grater s'tièt**e...
il soulève sa casquette [pour lui gratter sa tête] pour gratter sa tête

*****

(trouvé en français à l'écrit) : **j'ai fait un trou pour les fils passer dedans**

>>> interférence ou continuum ou système commun ?

Des données vers la théorie (6)

mais le corpus donne aussi ceci :

I falloait nin profiter **pour li mette sin nez** din sin brin !
Il fallait en profiter pour lui mettre son nez dans son caca

Pi si q cho vient à t-ête con.nu, vite o dessaque un lampisse **pour li foaire porter ch'capieu**.
et si cela vient à être connu, vite on sort un lampiste pour lui faire porter le chapeau

Dusqu'à doù qu'il iront **pour nous foaire ainmer l'an 2000**, conme i dit'te ?
jusqu'où iront-ils pour nous faire aimer l'an 2000, comme ils disent ?

Conclusions