]

Introduction à la morphologie à deux niveaux

Benoit Crabbe

2012-2013

Outline

- Morphologie exionnelle et derivationelle: qq faits
- Modelisation informelle (introduces rewriting rules)
- Formalisation par transducteurs a etats nis:
  - Automates à états finis (recap:déterminisme)
  - Relations régulières et transducteurs
  - Principales opérations (union, intersection, composition)
  - Expressions régulières pour la morphologie
  - Ambiguité et transducteurs p-séquentiels

= *how to properly formalize a language problem, (@see Sohn 2008)*

Plan

1. Fragment de morphologie du francais
2. Modelisation (Automates a etats nis)
3. Relations regulieres
4. Transducteurs
5. Langages d'expressions regulieres pour la morphologie

Pourquoi l'analyse morphologique ?

- Normalisation des mots (lemmatisation) : utile en RI
- Normalisation des mots (proprietes exionnelles) : exemple de la resolution d'anaphores.
- Description intensionelles du dictionnaire, possibilite de le compresser ; examiner les processus de creation des mots, neologismes

Variation surfacique et morphologique

- Variations surfacique des mots
  - Télé ; Télévision ; T.V. (différents mots, non traité ici)
  - télévision, télévisions, télévisé, télévisuel, télévisuels...
  - sport, sports, sportif, sportifs,...
- Comment faire le match ?
  - Convertir les formes observées (en surface) à des formes basiques (canoniques)
  - Stemming ou analyse morphologique

Approche énumérative

- Creer un dictionnaire avec toutes les formes de mots; associe les mots observes a leurs formes canoniques.
- Faisable (Le f).
- Probleme, on peut construire constamment des nouveaux mots par procedes reguliers
- Exemples : talibaniser, pipolisation, sarkosisme, cybercafe,depaternalisation...
- Il existe des langues comme le turc ou les mots sont agglutines, on y trouve des mots de 40 lettres qui ont valeur de phrases en francais. Enumerer le vocabulaire turc ne fait pas sens.
- L'approche par enumeration manque de generalite.

Éléments de morphologie du français

On distingue mots simples et mots composes

- **Mots simples :** approx chaine de caracs separes par des blancs
  - Radical simple
  - Radical + morphème (genre, nombre, personne, marqueur de dérivation)
  - Abbréviations

Morphème

= plus petit segment porteur de sens (= plus petite unite obtenue par segmentation du sens des enonces, = plus petite unite porteuse de sens.)

- **Mots composés:** contiennent blancs et tirets:
  - Unité conceptuellement autonome
  - Sens $\neq$ somme de ses parties (croque-monsieur, perce-oreille)

Mots dérivés
Principes de formation

- **Formation :** pre xe ou su xe
- Deux modes de formation:
  - Pas de transfert catégoriel : Agréable → désagréable
  - Transfert catégoriel : blanc → blanchir

Lemmes vs racines

Type de derivation	Lemme	Derive
su xation	Lire	lisible
pre xation	lisible	illisible
su xation	illisible	illisiblement

On a une **relation** de derivation entre lemmes et derives. On distingue le lemme de la racine morphologique (sans la exion)

Affixes

- A xation: noyau morphologique+satellites
- A xes exionnels (exion nominale, verbale, adjectival)
- A xes derivationnels (pre xe, ex. anti-; su xe, ex - ment)
- Comment trouver des a xes (en partic. derivationnels), test de commutation:

Psycho	motricite	vs	??de	marrer
	therapeute			pêcher

(le sens change du tout au tout)

Préfixation

- Autonomie, distribution, semantique, transfert de categorie grammaticale, combinatoire morphologique
- Exercice (observations):

  Progouvernemental, antiraciste, bimensuel, interurbain, préhistorique, surnuméraire, illisible, déshabiller, recommencer

- Consignes:
  - Entourer le préfixe de chaque mot
  - Pour chacun d'entre eux, identifier le trait sémantique qui le caractérise parmi les traits suivants : *Privatif, temporel, intensif, réitératif, spatial, quantitatif, oppositif, approbateur*

Suffixation

- Autonomie, distribution, semantique, transfert de categorie grammaticale, combinatoire morphologique
- Robert en recense environ 200
- Typologie (changement de categories):
- Su xe nominal
  - forme verbale → nominale : ronger, rongeur
  - forme adjectivale → nominale : rouge, rougeur
- Su xe adjectival
  - forme nominale → adjectivale: nation, national
  - forme verbale → adjectivale: manger, mangeable
- Su xe verbal
  - forme nominale → verbale: clou, clouer
  - forme adjectivale → verbale: facile, faciliter
- Su xe adverbial
  - forme adjectivale → adverbiale: facile, facilement
- Su xe diminutif/pejoratif
  - forme nominale → nominale: roi, roitelet
  - forme adjectivale → adjectivale : long, longuet
  - forme verbale → verbale: voler, voleter

Suffixation
Exercices

Machinerie, faciliter, rapidement, clouer, mangeable, rougeâtre, rondelet, mordiller, frottement, radicalisme, national, voleter, commercialisation, monétarisme, responsables, incapables, inhabituel, fortement, production, enseigne, agriculteur, chou-fleur.

- Entourer le su xe
- Donner le terme de base (lemme)
- Donner la categorie du lemme, du mot derive
- Indiquer les termes pour lesquels il y a transfert de categorie

Variantes surfaciques

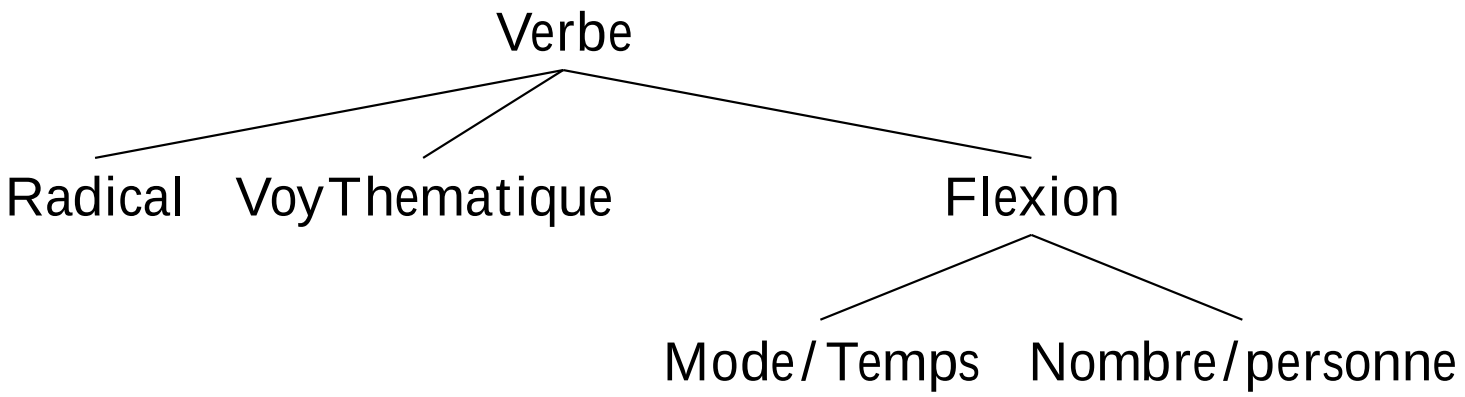
Parfois le processus de derivation entraine certaines irregularites surfaciques:

- fort ; fort-e-ment
- fort ; fort-i er
- facile ; facil-it-er
- facile; facilement

Morphologie flexionnelle

- Pas de nouveaux mots ; variantes du même mot
- ex : temps, mode, personne, genre, nombre
- En n de mot apres les su xes derivationnels

Flexion verbale



- Radical (porteur du sens lexical)
- Voyelle thematique : correspond a l'allongement du radical aim**er** / **nir**
  - Conjugaisons thématiques: (aimer; finir)
  - Conjugaisons athématiques: (vendre; conclure)
- Morphemes exionnels: se succedent dans un ordre contraint: temps / mode : je vend**rai**; personne / nombre je vendr**ai**
- On suppose un morpheme nul
- ...(il existe des irregularites)

Irrégularités apparentes (de surface)

- je seme, tu seme . . . nous semons, vous semez. . .
- Irregulier ?
- pas tout a fait : alternance vocalique
  - Une syllabe ouverte se termine par une voyelle
  - Une syllabe fermée se termine par une consonne
- **Règle phonologique générale :** Une voyelle en syllabe fermee doit être ouverte (o ouvert,e,a,e ouvert), une voyelle en syllabe ouverte doit être fermee (i,u,o,e)
- Ici, en syllabe fermee (il seme) on a voyelle ouverte, en syllabe ouverte (se-mons) on a une voyelle fermee.
- tu bats / nous bat-tons ; je vis,tu vis.. nous vivons
- allomorphes : je sais , nous savons; (fully irregular) je vais , nous irons..

Modélisation de la morphologie flexionelle du verbe

Morphologie à deux niveaux

Buts de la manoeuvre

- On suppose une forme de surface a traduire par une forme analytique
- Gerer les e ets phonologiques (doublement de lettres, insertion de voyelles. . .)
- Gerer les exceptions pathologiques

Exemples:

- Aimer:
  Forme analytique : aimer+V +ind +ons+1+pl
  Forme surfacique : aim ϵ ons
- Cheval (mais des carnavals):
  Forme analytique : cheval+N +masc +pl
  $l \rightarrow$ u/[a]_[+x]
  Forme surfacique : chevau ϵ x
- semer:
  Forme analytique : semer+V +ind +1+sg
  $e \rightarrow$ e/[\$$C$]_[\$$C$+]
  Forme surfacique : sem ϵ e

Plan

1. Fragment de morphologie du francais
2. Modelisation (Automates a etats nis)
3. Relations regulieres
4. Transducteurs
5. Langages d'expressions regulieres pour la morphologie

Automate a etats nis

Une bande de lecture

Une machine qui la lit

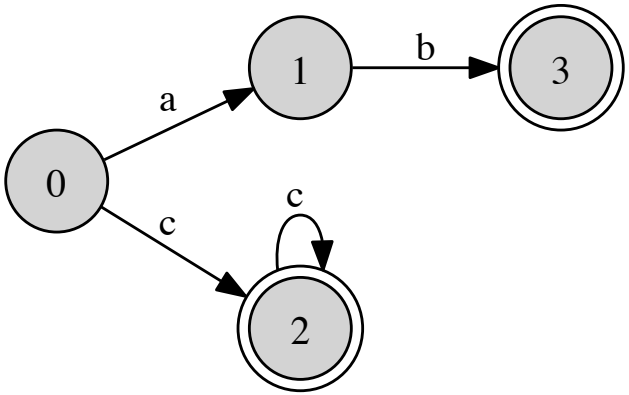
Acceptation

- Etant donne un automate deterministe $M = \langle \quad, Q, i, F, \delta \rangle$,
- M **accepte** x ssi il existe $q \in F$ tel que $(\epsilon, i, x) \Rightarrow_* (x, q, \epsilon)$

Langage accepté par M

Le langage $L(M)$ accepte par M est l'ensemble de toutes les sequences de symboles acceptees par M

Exemple



- $L(M)$ = $(ab|c+)$
- Soit x = ab ;
- Derivation:

$$(0, 0, ab) \Rightarrow (a, 1, b) \Rightarrow (ab, 3, \epsilon)$$

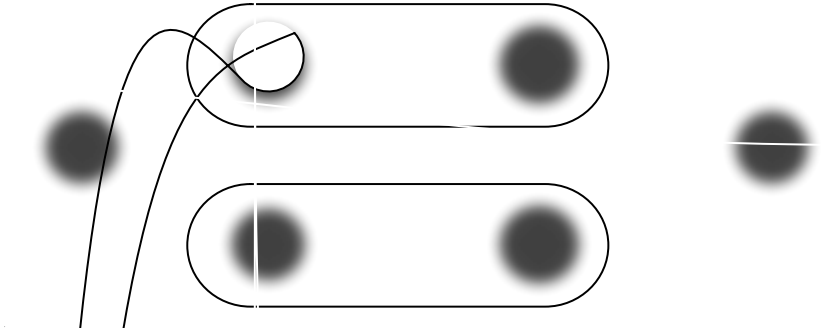
ce qui est une acceptation car ab = x et $3 \in F$

Automates finis et expressions régulières

- Operations elementaires sur les automates nis non deterministes: (pour construire des automates plus complexes a partir d'automates simples)
- **Union** de M_A et de M_A, automate resultant:
  - $i = q_0$ (nouvel état initial)
  - $F = \{q_f\}$ (nouvel état final unique)
  - $\epsilon-$transitions :

$$\Delta = \Delta_A \cup \Delta_B \begin{array}{ll} \cup & \{\delta((q_0, \epsilon), q_z) | q_z \in \ i_A, i_B\} \\ \cup & \{\delta((q_z, \epsilon), q_f) | q_z \in F_A \cup F_B\} \end{array}$$

- Illustration:



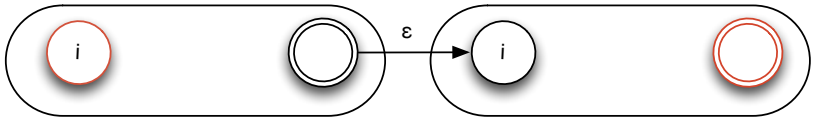
Fermeture (étoile) de Kleene

- **Fermeture de Kleene** pour M_A, automate resultant:
  - $i_* = q_0$ (nouvel état initial)
  - $F_* = \{q_f\}$ (nouvel état final unique)
  - $\epsilon-$transitions :

$$\Delta_* = \Delta \begin{array}{ll} \cup & \{\delta((q_0,\epsilon),i_A),\delta((q_0,\epsilon),q_f)\} \\ \cup & \{\delta((q_z,\epsilon),q_f)|q_z \in F_A\} \\ \cup & \{\delta((q_z,\epsilon),i_A)|q_z \in F_A\} \end{array}$$

Concaténation

- **Concaténation** de M_A et de M_A, automate resultant:
  - $i = i_A$
  - $F = F_B$
  - $\Delta = \Delta_A \cup \Delta_B \cup \{\delta((q_i, \epsilon), q_f) | q_i \in F_A, q_f = i_B\}$



Expressions régulières
=Ajout d'une syntaxe

- ϵ est une expression reguliere
- a est une expression reguliere ssi $a \in$
- $A|B$ est une expression reguliere ssi A est une expression reguliere et B est une expression reguliere [union des 2 automates]
- AB est une expression reguliere ssi A est une expression reguliere et B est une expression reguliere [concatenation des 2 automates]
- $A*$ est une expression reguliere ssi A est une expression reguliere [kleene star de l'automate]

Sucrage syntaxique (qq exemples)

- Parenthésage : (A)
- (au moins une occurrence) : $A+ =_{\text{def}} a \quad a$
- Any symbol: $? =_{\text{def}} (\sigma_1|\ldots|\sigma_n) \qquad \forall \sigma_i \in \Sigma$
- Appartenance : $\$A =_{\text{def}} (? \quad A? \ )$
  (exemple : $\$(a|b)$ matche toutes les séquences qui contiennent un a ou un b)

Les REs forment un langage de plus haut niveau qui permet d'encoder des automates très complexes sans avoir à tripatouiller des noeuds et des transitions. Pas créé pour faire de la linguistique ceci dit.

Intersection

Algorithme pour le cas déterministe

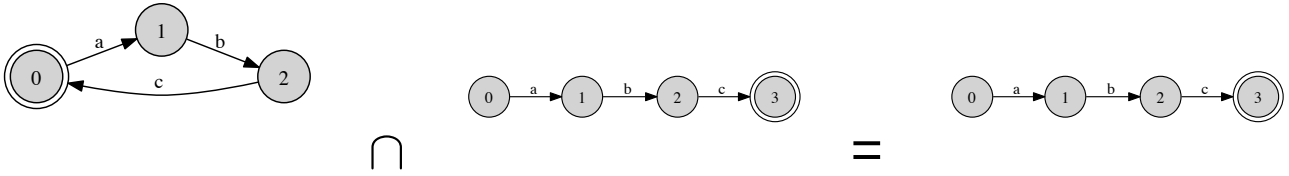
Opération quadratique !

L'intersection est une operation generalement rendue inaccessible au grand public (compilateurs de regexes type Perl,java, python) car quadratique.

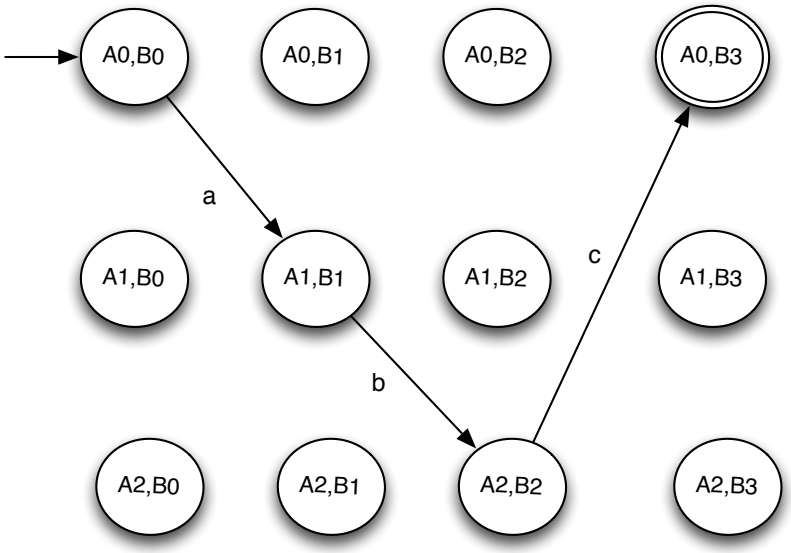
- L'algorithme d'intersection de deux automates A et B produit un automate C tel que $L(C) = L(A) \cap L(B)$
- On construit C comme suit:
  - $Q = Q_A \times Q_B$
  - $i = (i_A, i_B)$
  - $\Delta = \{((q_{A_i}, q_{B_i}), \sigma, (q_{A_j}, q_{B_j}))|(q_{A_i}, \sigma, q_{A_j}) \in \Delta_A \wedge (q_{B_i}, \sigma, q_{B_j}) \in \Delta_B\}$
  - $F = \{(q_A, q_B)|q_A \in F_A \wedge q_B \in F_B\}$
- **Note:** en cas d'echec (par exemple pas d'etat nal), l'intersection renvoie un automate a un etat initial- nal pour representer le langage vide.

Intersection
Exemple

- Soient A et B:

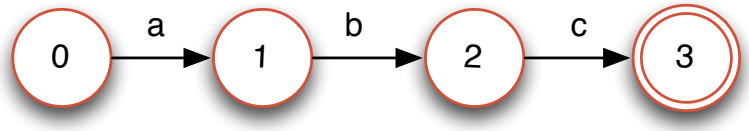


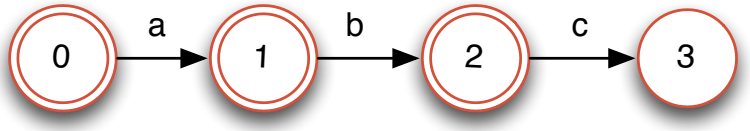
- Detail :



Complémentation
Cas déterministe

- La complementation d'un automate deterministe A produit un automate A tel que $L(A)$ = $* - L(A)$
- La procedure consiste a intervertir les etats naux
  - $F = Q - F$
- Exemple:





Relations régulières et transducteurs

- On a vu qu'a partir d'une notion simple d'automate et de quelques operations bien de nies dessus (union,concatenation,fermeture,determinisation, minimisation) on peut construire un langage relativement riche d'expressions regulieres
- On propose d'envisager le même procede pour le langage en utilisant des transducteurs qui donne une certaine expressivite pour exprimer les regles dites phonologiques :
  - Machine à deux bandes (lecture,écriture)
  - Elle lit sur l'une et écrit sur l'autre.

Plan

1. Fragment de morphologie du francais
2. Modelisation (Automates a etats nis)
3. Relations regulieres
4. Transducteurs
5. Langages d'expressions regulieres pour la morphologie

Relations régulières

- Les relations regulieres encodent une relation entre deux langages (lecture,ecriture)
- Soit $_1$, $_2$ les alphabets d'entree et de sortie, un ensemble de relations regulieres se de nit comme suit:
  - $\emptyset$ est une relation régulière
  - $\{(x, y)\}$ est une relation régulière ssi $(x, y) \in \Sigma_1 \times \Sigma_2$
  - (Concaténation) AB est la relation régulière $\{(x_1x_2, y_1y_2)|(x_1, y_1) \in A; (x_2, y_2) \in B\}$ ssi A est une relation régulière et B est une relation régulière
  - (Union) $A \cup B$ est une relation régulière ssi A est une relation régulière et B est une relation régulière $A \cup B = \{(x, y)|(x, y) \in A \vee (x, y) \in B\}$
  - (Fermeture) $A*$ est une relation régulière ssi A est une relation régulière

Composition de relations régulières

- (Concatenation) AB est la relation reguliere $\{(x_1x_2, y_1y_2)|(x_1, y_1) \in A; (x_2, y_2) \in B\}$ ssi A est une relation reguliere et B est une relation reguliere
- (Union) $A \cup B$ est une relation reguliere ssi A est une relation reguliere et B est une relation reguliere
  $A \cup B = \{(x, y)|(x, y) \in A \vee (x, y) \in B\}$
- Intersection (comme automates, consider symbols as pairs, non close si ϵ transitions!)
  Idee $(a^n b* : c^n) \cap (a * b^n : c^n)$ produit $(a^n b^n : c^n)$ qui n'est pas regulier.
- Composition de deux relations (new stu):

$$A \circ B = \{(x, z)|\exists y(x, y) \in A \wedge (y, z) \in B\}$$

Plan

1. Fragment de morphologie du francais
2. Modelisation (Automates a etats nis)
3. Relations regulieres
4. Transducteurs
5. Langages d'expressions regulieres pour la morphologie

Transducteur à états finis

Définition

Un transducteur a etats nis est un tuple $\langle \ _1, \ _2, Q, i, F, E \rangle$ ou :

- $_1$ est un ensemble ni de symboles (alphabet d'entree)
- $_2$ est un ensemble ni de symboles (alphabet de sortie)
- Q est un ensemble ni d'etats
- $i \in Q$ est l'etat initial
- $F \subseteq Q$ est un ensemble d'etats naux
- $E \subseteq Q \times (\ _1^* \times \ _2^*) \times Q$ est la relation de transition

Propriété désirable

Un transducteur est **séquentiel** si aucun etat n'a deux arcs sortants avec le même symbole d'entree. (automate deterministe en entree)

Dérivation

Run d'un transducteur

- Une con guration M est un triplet (x, q, y) ou $x \in \ {}^*_1$, $y \in \ {}^*_2$ et $q \in Q$.
  où x est la séquence lue, y la séquence déjà écrite
- (Mouvement) La con guration (x, q, y) **dérive immédiatement** la con guration (x', q, y') (note $(x, q, y) \Rightarrow_1 (x', q', y')$) ssi:
  - Il existe un couple de symboles $(\sigma_1, \sigma_2) \in \Sigma_1 \times \Sigma_2$ tel que $x' = x\sigma_1$ et tel que $y' = y\sigma_2$
  - $(q, \sigma_1, \sigma_2, q') \in E$
- La con guration M derive la con guration N (note $M \Rightarrow_* N$) ssi il existe une sequence $S_0 \Rightarrow_1 S_1 \Rightarrow_1 \ldots \Rightarrow_1 S_n$ telle que S_0 = M et S_n = N

Acceptation

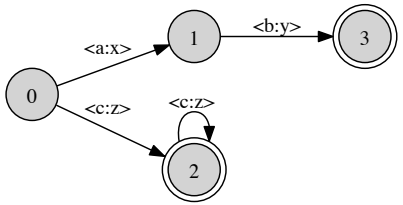
- La **configuration initiale** M_I d'un transducteur A est la con guration $(\epsilon, i_A, \epsilon)$
- La **configuration finale** M_F d'un transducteur A est la con guration $(x\epsilon, q_f, y\epsilon)$ telle que $q_f \in F_A$.
- A **accepte** x ssi $M_I \Rightarrow_* M_F$

Transducteurs à états finis

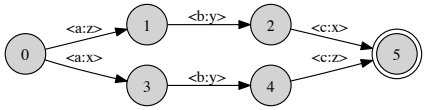
Exemples

Le transducteur sequentiel est la contrepartie de l'automate deterministe: pas deux transitions sortantes avec le même label sur l'arc. Ceci dit la sequentialite ne veut pas dire que l'output est unique (cas du transducteur in niment ambigu)

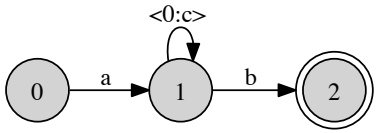
- Sequentiel



- Non sequentiel



- Un autre in niment ambigu (non sequentiel)



Opérations sur les transducteurs
Composition

- Concatenation (comme automates)
- Union (comme automates)
- Determinisation, Minimisation ($\approx$ automates, consider symbols as pairs, but beware. . . sequentiality, functionality)
- Algorithme de composition des transducteurs A et B:
  - $Q = \{(q_a, q_b)|Q_A \times Q_B\}$ (produit)
  - $i = i_A \times i_B$ (produit)
  - $F = F_A \times F_B$ (produit)
  - $E = \{((q_{A_i}, q_{B_i}), x, z, q_{A_j}, q_{B_j}) \mid \exists y[(q_{A_i}, x, y, q_{A_j}) \in A \wedge (q_{B_i}, y, z, q_{B_j}) \in B]\}$

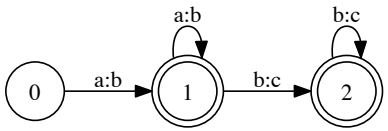
  Une vraie implementation supprime les arcs et etats non atteignables depuis le nouvel etat initial

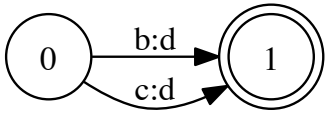
Complexité

Aie ! C'est quadratique ! (produits initiaux)

Composition (exercice)

- Soit les transducteurs t_1 et t_2:





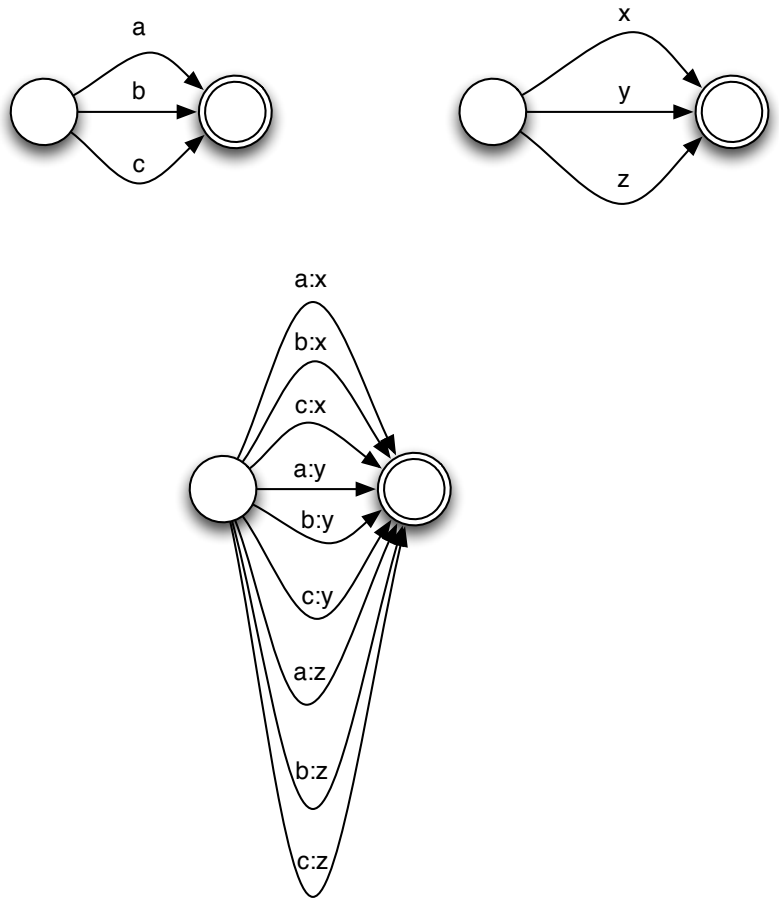
- Calculer $t_1 \circ t_2$ (minimisation optionnelle):
  - $Q_c = Q_A \times Q_B = \{(0_1, 0_2), (0_1, 1_2), (1_1, 0_2), (1_1, 1_2), (2_1, 0_2), (2_1, 1_2)\}$
  - $I_c = \{(0_1, 0_2)\}$
  - $F_c = \{(1_1, 1_2), (2_1, 1_2)\}$
  - $\delta_C = \{(q(0_1, 0_2, a, d) \rightarrow (1_1, 1_2)), (q(1_1, 0_2, b, d) \rightarrow (2_1, 1_2)), (q(1_1, 0_2, a, d) \rightarrow (1_1, 1_2)), (q(2_1, 0_2, b, d) \rightarrow (2_1, 1_2))\}$

Des transducteurs aux automates

et vice versa

- **Inversion:** L'inversion d'un transducteur A est construite comme suit:
  - $E = \{(q, y, x, q')|(q, x, y, q') \in E_A\}$
- **Projection**: L'automate a etats nis M resultant de la projection d'un transducteur A est construit comme suit:
  - On recopie tout sauf les transitions qu'on construit comme suit
  - $\Delta = \{((q, x), q')|(q, x, y, q') \in E_A\}$
- **Produit croisé (cartésien):** Soit $A = \langle Q_A, \quad _A, i_A, F_A, \quad _A \rangle$ et $B = \langle Q_B, \quad _B, i_B, F_B, \quad _B \rangle$ deux **automates finis**, le transducteur resultant de leur produit croise est construit comme suit:
  - $Q = Q_A \times Q_B$
  - $\Sigma_1 = \Sigma_A$
  - $\Sigma_2 = \Sigma_B$
  - $i = (i_A, i_B)$
  - $F = F_A \times F_B$
  - $E = \{((q_{A_i}, q_{B_i}), x, y, (q_{A_j}, q_{B_j}))|(q_{A_i}, x, q_{A_j}) \in \quad _A \wedge (q_{B_i}, x, q_{B_j}) \in \quad _B\}$

Produit cartésien (illustré)



Utilise en pratique pour faire correspondre les formes observees aux formes analytiques

Plan

1. Fragment de morphologie du francais
2. Modelisation (Automates a etats nis)
3. Relations regulieres
4. Transducteurs
5. Langages d'expressions regulieres pour la morphologie

Rappel général

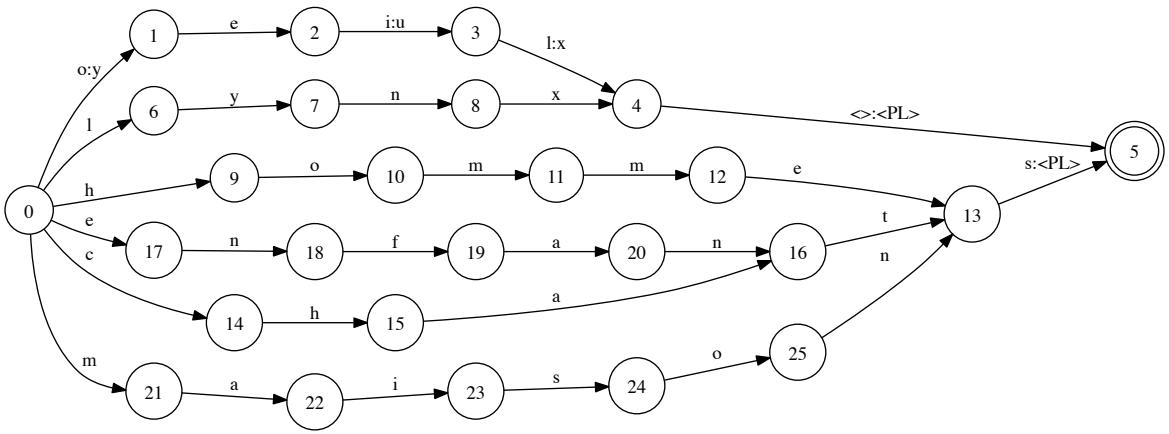
- Cas reguliers : aime : j'aim+e, tu aim+es, il aim+e, nous aim+ons, vous aim+ez, ils aim+ent
- Cas semi reguliers:

  Forme analytique : semer+V +ind +1+sg

  $e \rightarrow$ e/[\$$C$]_[\$$C$+]

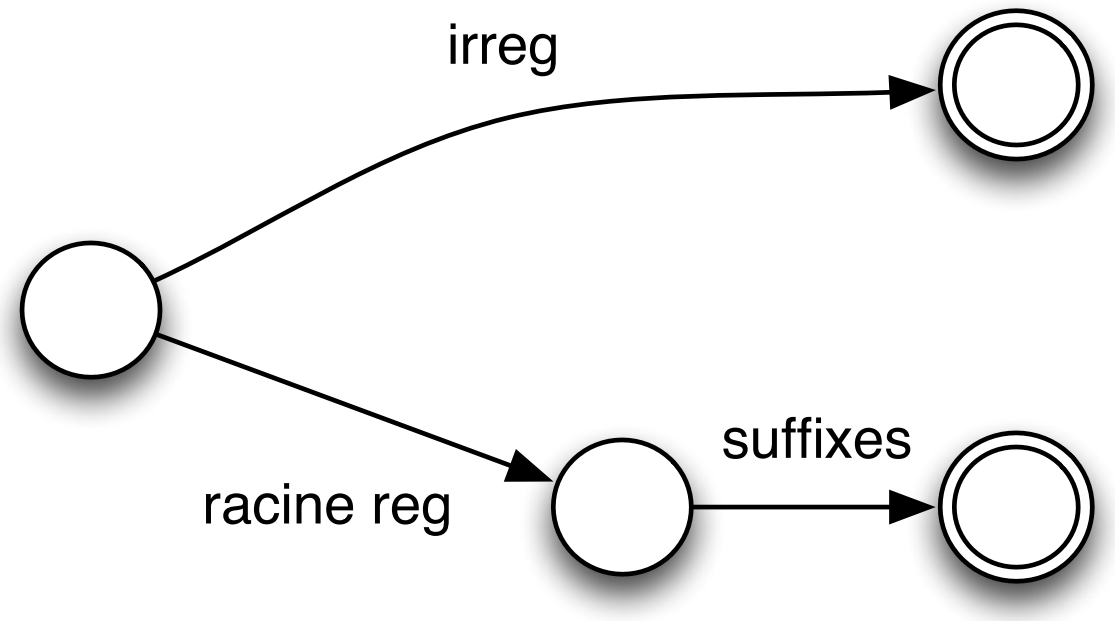
  Forme surfacique : sem ϵ e
- Cas irreguliers (aller): je vais nous allons, j'irai...

Idée générale de la formalisation



- On constuit un transducteur pour les racines (par produit cartesien) (1)
- On le concatene a un transducteur pour les terminaisons regulieres (2)
- On disjoint le tout avec un transducteur qui encode en dur les irreguliers (3)
- Reste les regles orthographiques : composition sur le transducteur regulier (1) + (2) avant de disjoindre

Schéma et exercice



o Règles orthographiques

- Representer les etapes de construction d'un transducteur pour analyser les verbes suivants: *aimer*, *garder*, *aller* a l'indicatif present
- Gerer les cas: *manger*, *ranger*

Règles orthographiques

- Insertion de e entre g et o
- Syntaxe de \linguiste" : $\epsilon \rightarrow e/g_o$
- Solution possible : $(g : g)(\epsilon : e)(o : o)$
- **En pratique :** (solutions bcp plus sioux) tient compte de la structure syllabique (pour eviter la surgeneration) plus complications liees a des cas particuliers de reduplication de symboles
- Les regles orthographiques (sont intersectees) ou composees entre elles et composees avec les automates de base
- Un langage d'expressions regulieres pour la morphologie donne une syntaxe aux regles orthographiques qui est traduite sous forme de transducteurs
- Exo pour la route (morpho derivationnelle): $classer$, $surclasser$, $classement$,$surclassement$. (la forme analytique indique la categorie du mot ex: +N, +ADV, +V+inf)