

Dictionnaires électroniques

Sébastien Paumier

`paumier@univ-mlv.fr`

Définition d'un dictionnaire classique

« Recueil des mots ou d'une catégorie de mots d'une langue, généralement rangés par ordre alphabétique (...) et expliqués dans la même langue ou traduits dans une autre » (Lexis, 1975)

Définition d'un dictionnaire électronique

Représentation formelle d'un lexique, qui associe à chaque forme fléchie son lemme ainsi que des informations grammaticales, flexionnelles et éventuellement sémantiques.

2 critères fondamentaux

Les dictionnaires électroniques doivent être:

- formels: peuvent être manipulés par des programmes
- exhaustifs: doivent couvrir 100% du lexique

Mots simples

Définition formelle:

séquence contiguë de lettres (définies par un alphabet)

Exemples:

point, .ADV+z1

poli, .A+z1:ms

poli, .N+z1:ms

poli, polir.V+z1:Kms

Définition des mots composés

Pas de définition claire et unanime, mais quelques critères:

- opacité sémantique (*cordon bleu* ≠ cordon qui est bleu)
- figement (*grand cru classé*, mais pas *immense cru fiché*)
- grammaticalisation (le *qu'en dira-t-on*)

Mots composés

Définition formelle:

séquence composée de plusieurs unités lexicales (mots simples, séparateurs, chiffres, etc)

Exemples:

cent soixante, .DET+Dnum:mp:fp

coup de sang, .N+NDN+z3:ms

coup de sifflet, .N+NDN:ms

coûte que coûte, .ADV+PCONJ+z1

Dictionnaires électroniques DELAF

`équilibristes,équilibriste.N+z1:mp:fp`



- `équilibristes`: forme fléchie
- `équilibriste`: lemme
- `N`: code grammatical
- `z1`: code sémantique pour « mot courant »
- `mp`: code flexionnel pour « masculin pluriel »
- `fp`: code flexionnel pour « féminin pluriel »

Flexion automatique

- Principe: générer les formes fléchies à partir des formes canoniques
- Méthode:
 - regrouper les mots en classes de flexion
 - décrire les classes de flexion
- Avantage: une seule forme à maintenir

Le DELAS

Associe à chaque lemme un code de flexion (catégorie grammaticale+numéro)

Exemples:

chaud ,N1

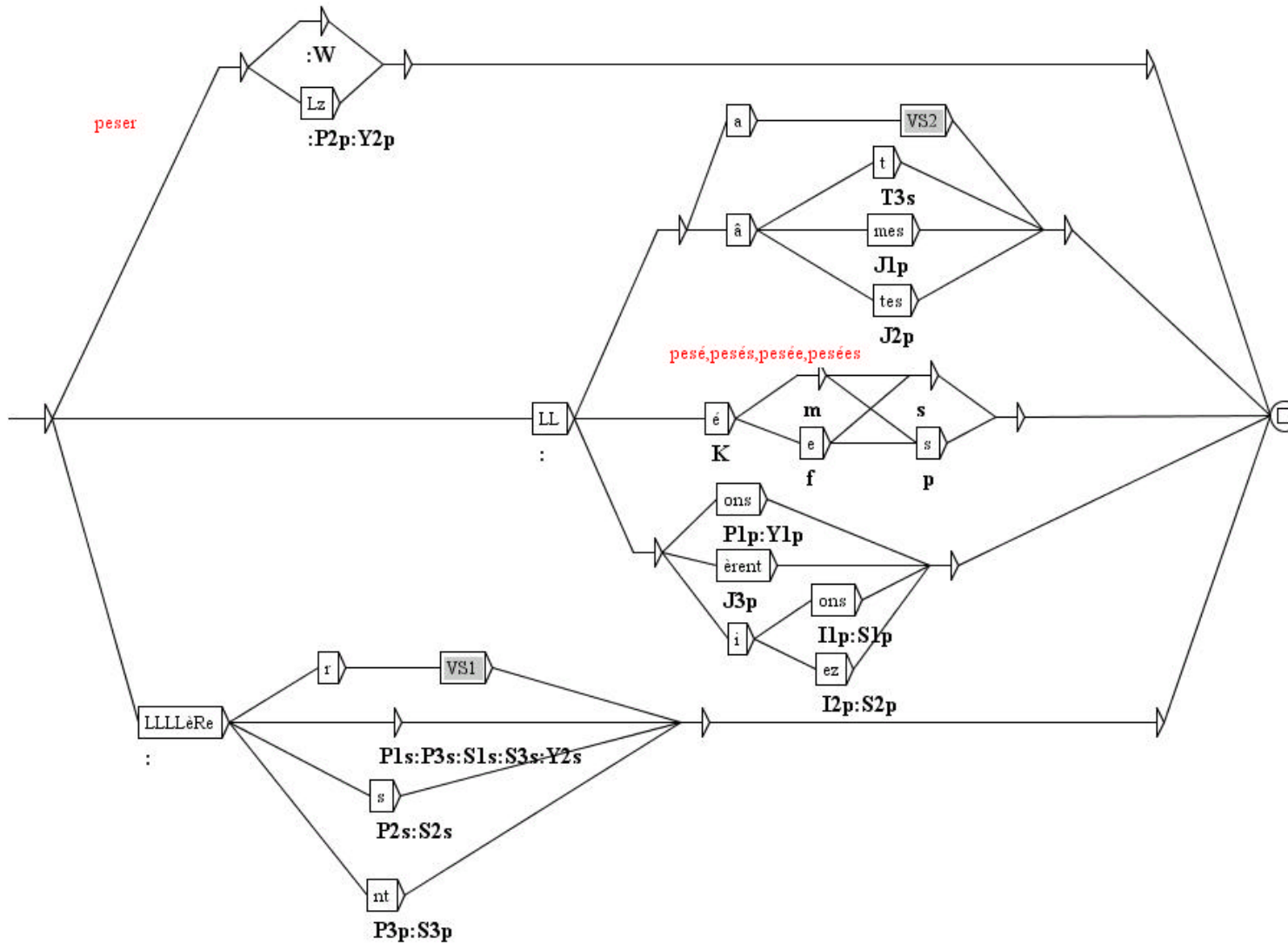
chaud ,A32

chauder ,V3

Grammaires de flexion

- Entrées: opérateurs à appliquer sur une pile
 - L: décaler le curseur vers la gauche (Left)
 - R: décaler le curseur vers la droite (Right)
 - C: décaler tout le sommet de pile en recopiant le caractère courant (Copy)
 - caractère quelconque: mettre le caractère dans la case courante et décaler le curseur vers la droite
- Sorties: codes flexionnels à concaténer

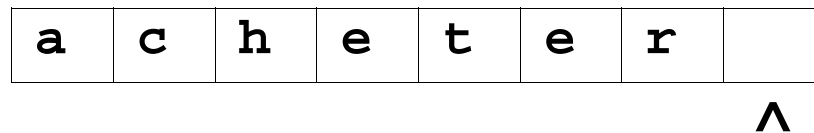
Exemple de grammaire



Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes/ :P2s:S2s

Étape 0: initialisation de la pile



Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 1: LLLLèRes

a	c	h	e	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 2: **LL**LLèRes

a	c	h	e	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 3: **LLL**LèRes

a	c	h	e	t	e	r	
---	---	----------	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 4: LLLLèRes

a	c	h	e	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 5: LLLLèRes

a	c	h	è	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 6: LLLLèRes

a	c	h	è	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 7: **LLLLèRes**

a	c	h	è	t	e	r	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter, V6
- Chemin de la grammaire V6: LLLLèRes / :P2s:S2s

Étape 8: **LLLLèRes**

a	c	h	è	t	e	s	
---	---	---	---	---	---	---	--

^

Exemple de flexion

- Entrée du DELAS: acheter ,V6
- Chemin de la grammaire V6: LLLLèRes/ :P2s:S2s

Étape 9: générer la ligne du DELAF

achètes ,acheter .V:P2s:S2s

Flexion des mots composés

- Plus difficile que les mots simples, car il faut savoir quelles parties fléchir :
 - **pomme** de terre
 - grand-**mère**
 - **contrôleur fiscal**
 - cessez-le-feu
- Adaptation au modèle de flexion des mots simples en cours (Savary)

Compression

Revuz (1990):

représentation d'un lexique



transducteur

Entrées: forme fléchies

Sorties: lemme+informations grammaticales, sémantiques et flexionnelles

Codage des entrées

chevaux de bois,cheval de bois.N:mp



Forme fléchie: chevaux de bois

Code de compression: 21 0 0.N:fp

Codage des entrées

chevaux de bois, cheval de bois.N:mp



Forme fléchie: chevaux de bois

Code de compression 21 0 0.N:fp

Enlever **2** lettres et ajouter **1**: chevaux ⇒ cheval

Codage des entrées

chevaux de bois,cheval de bois.N:mp



Forme fléchie: chevaux de bois

Code de compression: 21 0 0.N:mp

Enlever 0 lettre à de et 0 lettre à bois

Codage des entrées

Numérotation des codes de flexion:

0: .DET:fs

1: .PRO:fs

2: .N:fs

Regroupement des codes, pour qu'il n'y en ait qu'un par
forme fléchie:

la, .DET:fs

la, .PRO:fs ⇒ 3: .DET:fs, .PRO:fs, .N:fs

la, .N:fs

Construction du transducteur

chat, .N:ms

chaud, .A:ms

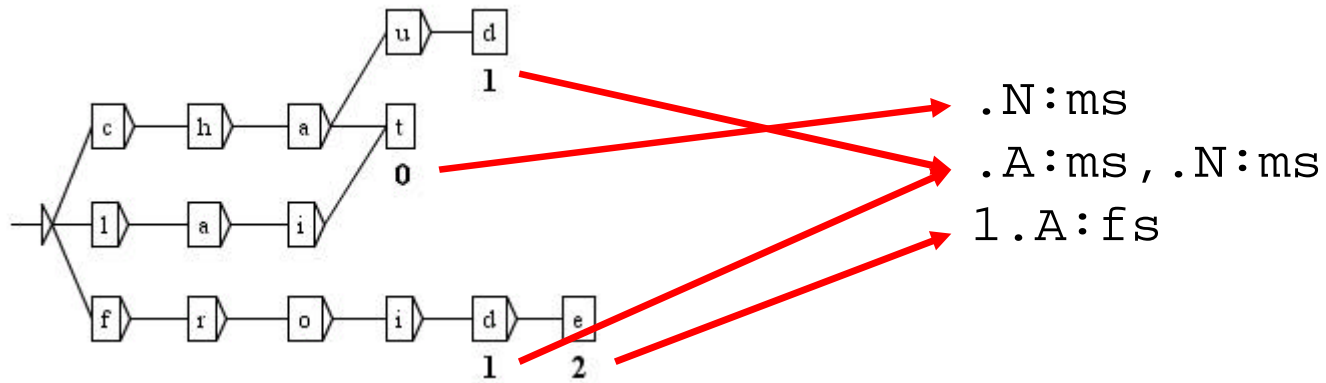
chaud, .N:ms

froid, .A:ms

froid, .N:ms

froide, froid.A:fs

lait, .N:ms

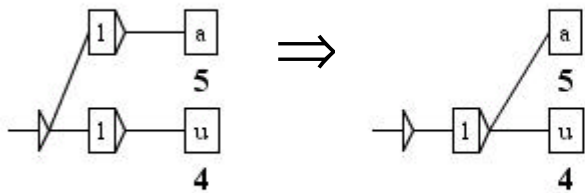


Minimisation

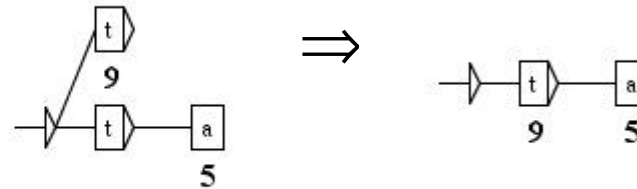
- Déterminisation possible car il n'y a qu'un code associé à chaque mot
- Minimisation: on fusionne les nœuds qui ont la même lettre d'entrée, sauf s'ils ont des lettres de sorties non vides et différentes

Fusion des nœuds

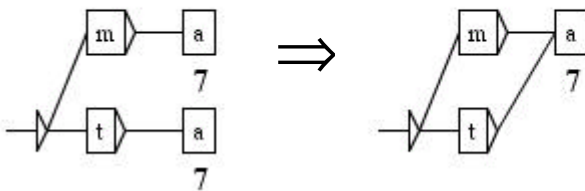
sortie vide/sortie vide



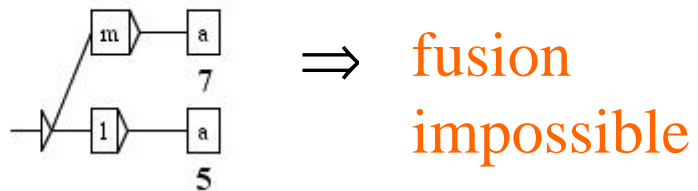
sortie vide/sortie



mêmes sorties



sorties différentes



Algorithmes de minimisation

- Méthode naïve
 - pas de restriction, mais gros calculs
- Algorithme de Revuz (utilisé dans Unitex)
 - minimisation linéaire des automates acycliques (rapide, mais gourmande en mémoire)
- Algorithme de Daciuk
 - construction semi-incrémentale d'un automate acyclique (économe en mémoire, plus lent)

Méthode naïve

1. construire un automate
2. l'inverser
3. le déterminer
4. le réinverser
5. le déterminer à nouveau

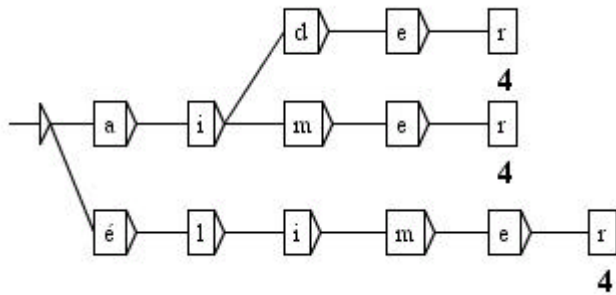
Algorithme de Revuz

Optimisation de l'algorithme d'Hopcroft pour les automates acycliques

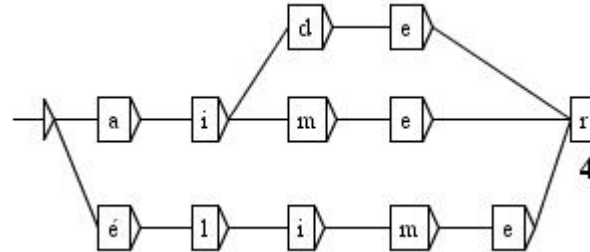
1. construire un arbre lexicographique (automate déterministe)
2. trier les nœuds de l'arbre par hauteur
3. fusionner, itérativement sur la hauteur, les nœuds de même hauteur, même contenu et mêmes transitions sortantes

Algorithme de Revuz

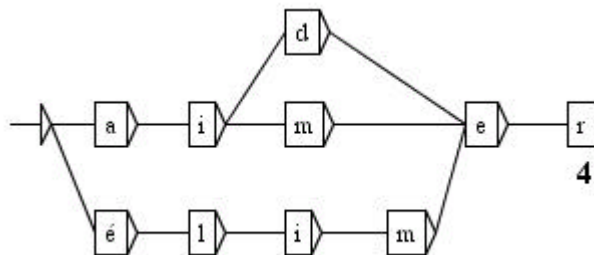
1) construction de l'arbre



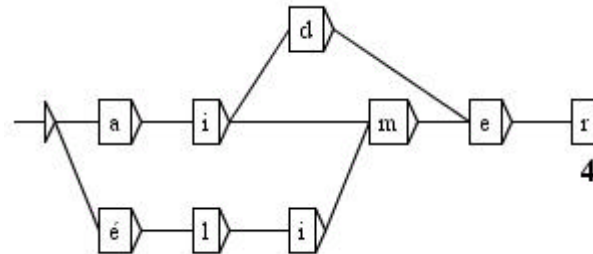
2) fusion des nœuds en r



3) fusion des nœuds en e



4) fusion des nœuds en m



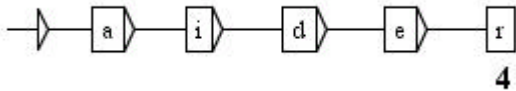
Algorithme de Daciuk

Principe: ajouter les mots les uns après les autres

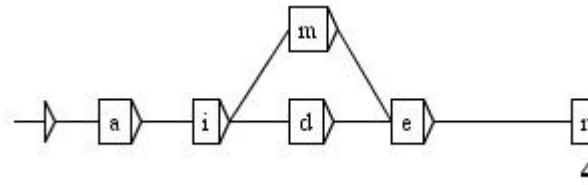
1. chercher les plus longs préfixes et suffixes ne se chevauchant pas, non ambigus et communs entre le mot et l'automate
2. ajouter ce qui reste du mot entre la fin du préfixe et le début du suffixe

Algorithme de Daciuk

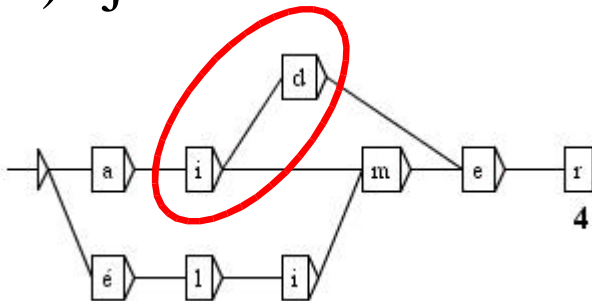
1) ajout de aider



2) ajout de aimer



3) ajout de éliminer



on ne prend pas le suffixe imer car le nœud en i a une transition vers d'autres nœuds que ceux du suffixe mer

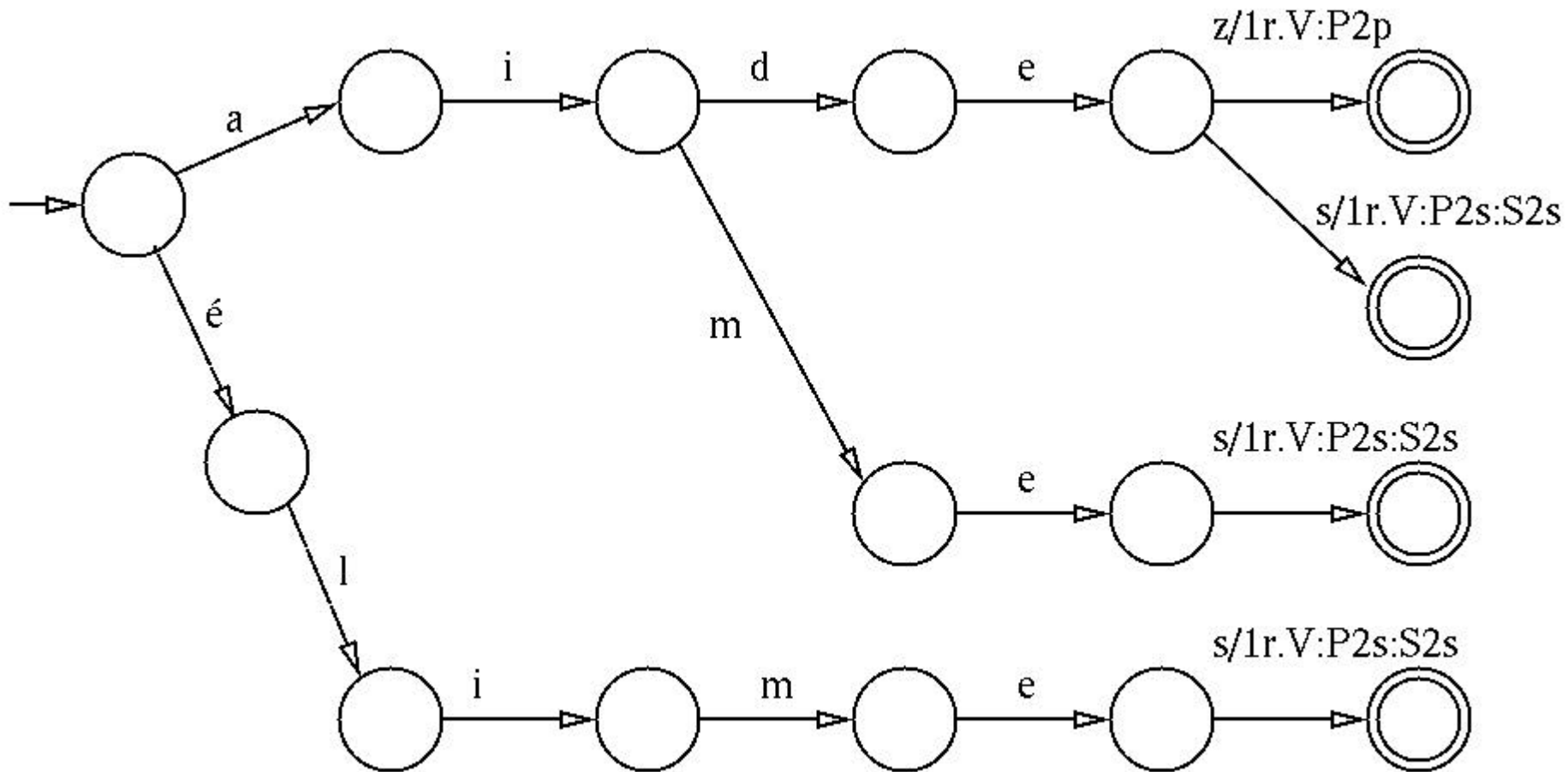
Autre méthode de compression

- coder les informations directement dans les sorties, sans passer par un numéro
- minimiser le transducteur (Mohri, Béal-Carton, Breslauer)

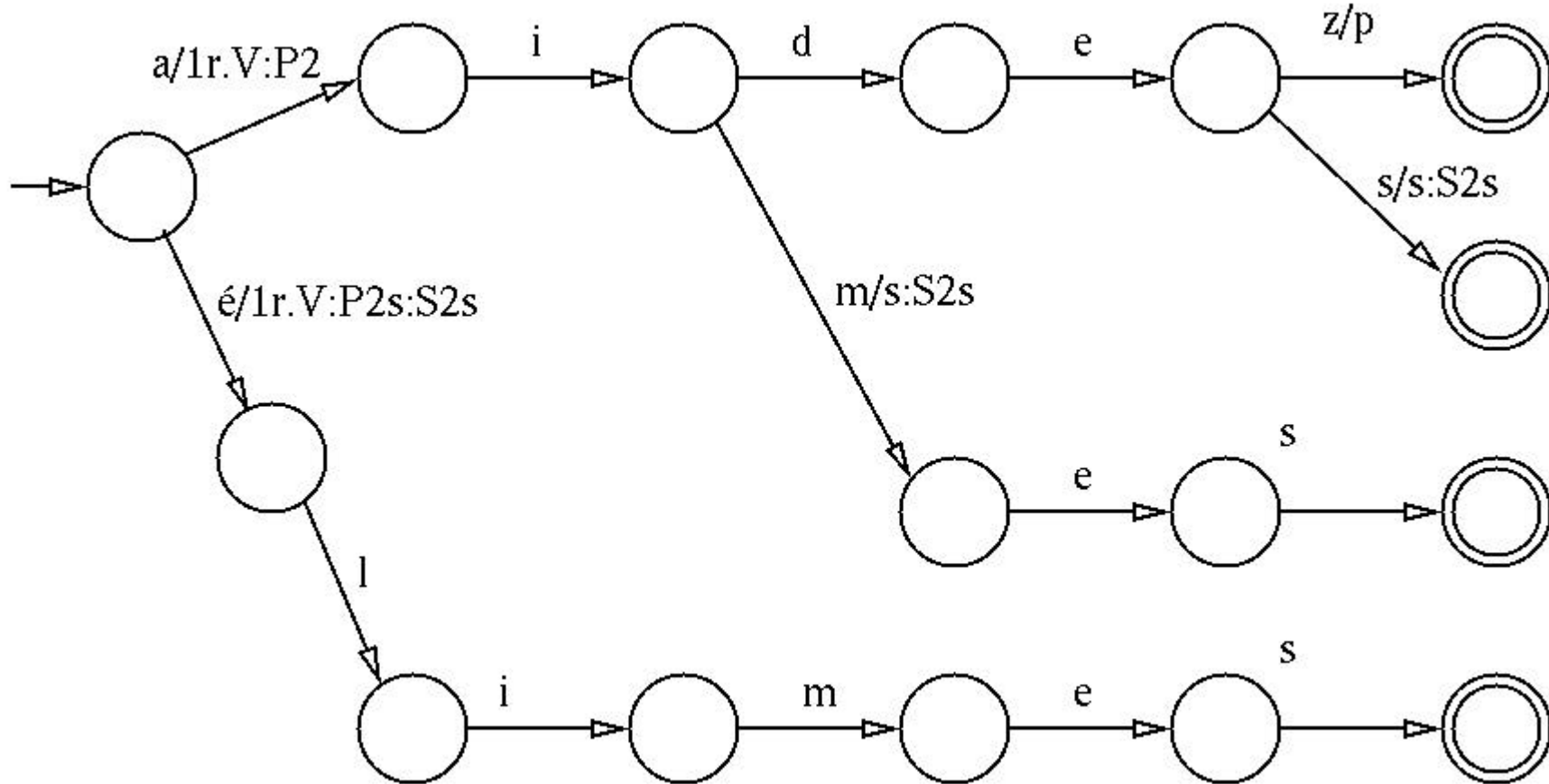
Minimisation

1. construire un transducteur déterministe en entrée
2. pousser les sorties le plus à gauche possible
3. minimiser le résultat en considérant chaque couple (*entrée, sortie*) comme un seul symbole

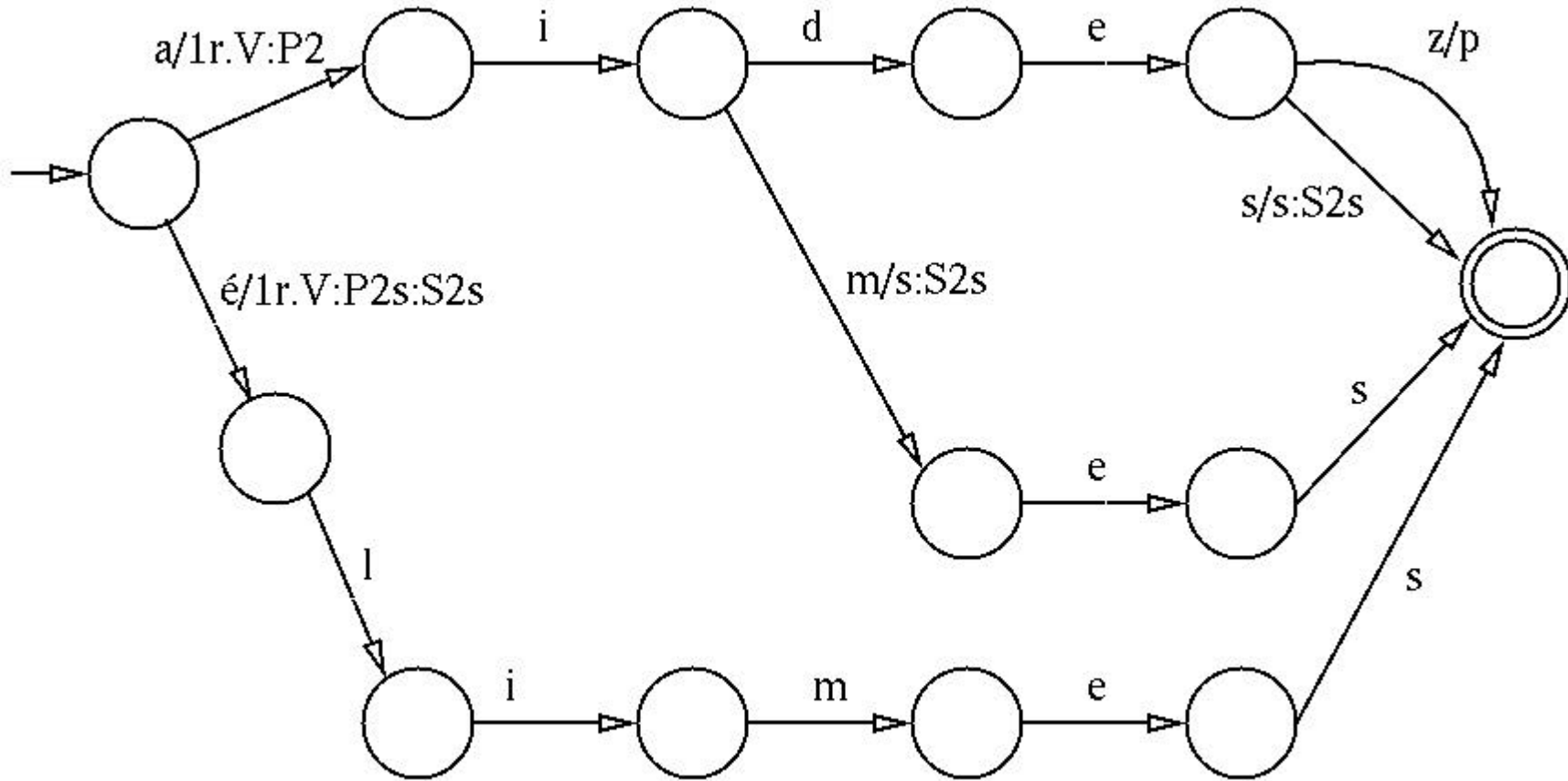
Transducteur déterministe sur les entrées



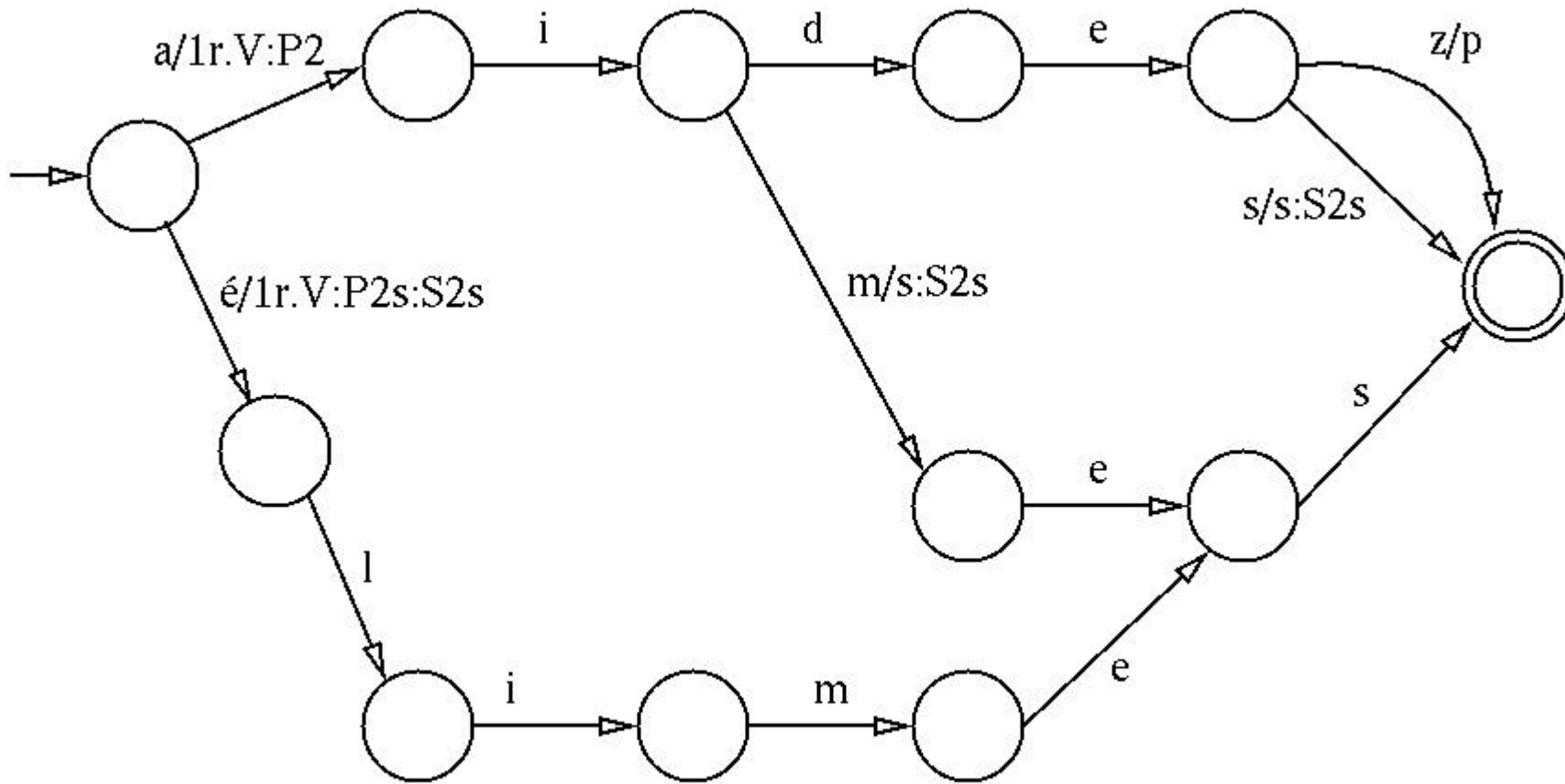
Poussée des sorties à gauche



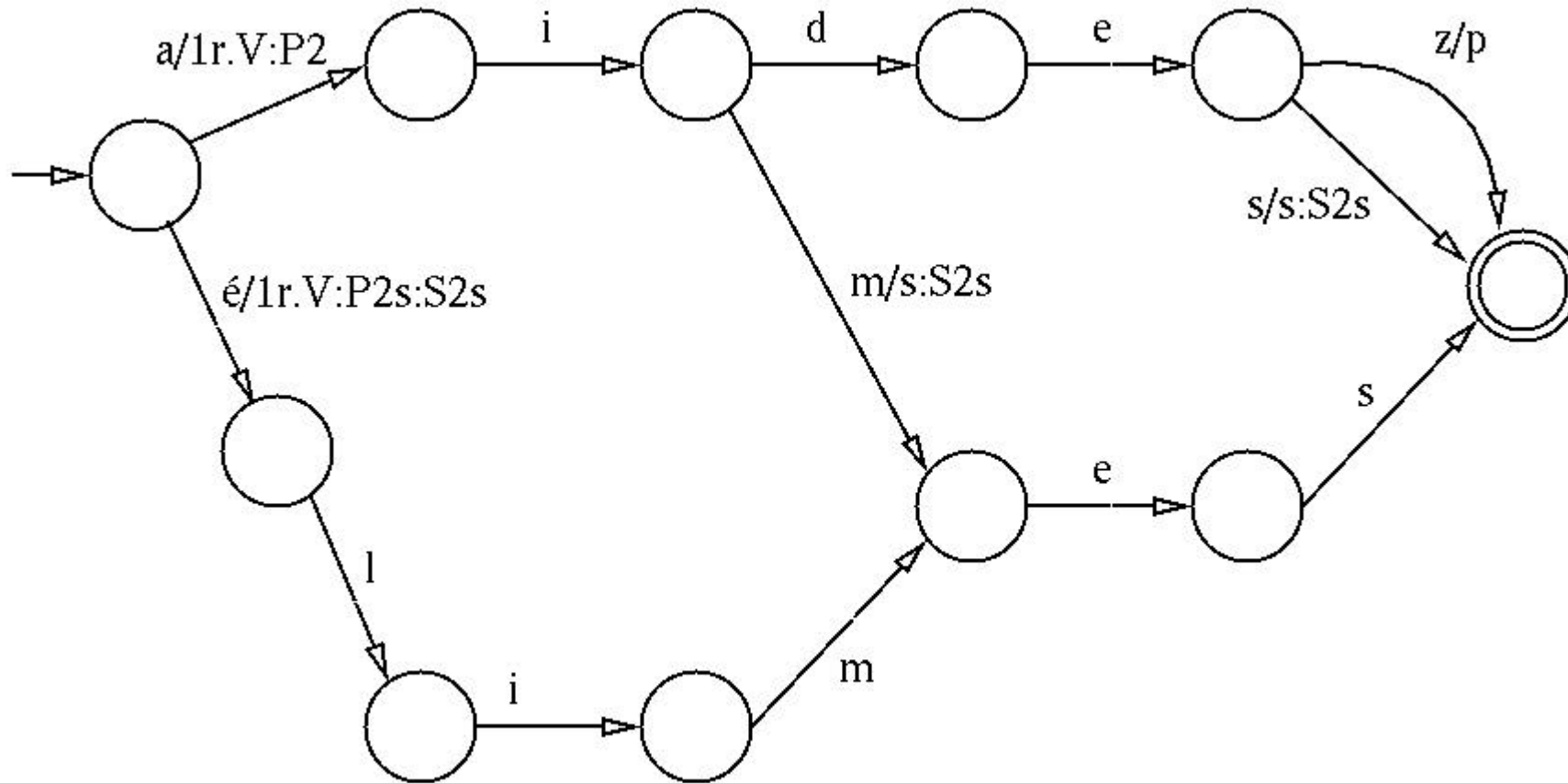
Fusion des nœuds de hauteur 1



Fusion des nœuds de hauteur 2



Fusion des nœuds de hauteur 3



Représentation d'un dictionnaire compressé

Dico.dic

le, .DET:ms

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms

Nombre de codes
présents dans le fichier

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Un code par ligne,
numérotation à partir de 0

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Taille du fichier
en octets

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

10000000 00000001

État non
final

Nombre de transitions
sortant de l'état

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Code du caractère étiquetant la
transition (006C: 1)

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Numéro du premier octet de
l'état pointé par la transition : 00000B

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

10000000 00000001

État non
final

Nombre de transitions
sortant de l'état

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Code du caractère étiquetant la
transition (0065: e)

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Numéro du premier octet de
l'état pointé par la transition : 000012



Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

00000000 00000000

État final

Nombre de transitions
sortant de l'état

Représentation d'un dictionnaire compressé

Dico.dic

⇒

Dico.inf

le, .DET:ms

0000000001

.DET:ms



Dico.bin

00 00 00 17 80 01 00 6C 00 00 0B 80 01 00 65 00 00 12 00 00 00 00

Numéro du code de compression
associé à cet état final

Avantages de cette représentation

- Format binaire facile à charger
- Parcours très rapide par adressage direct

Ordres de grandeur des DELAF

DELAF	Texte	Compressé	Taux
français	40,2 Mo	1,7 Mo	4,3 %
anglais	13,2 Mo	1,4 Mo	10,6 %
allemand	12,5 Mo	862 Ko	6,5 %
grec	83,8 Mo	4,6 Mo	5,4 %
italien	35,9 Mo	1,2 Mo	3,3 %
norvégien	23,0 Mo	1,1 Mo	4,9 %
portugais du Brésil	24,8 Mo	848 Ko	3,3 %
espagnol	35,4 Mo	1,3 Mo	3,8 %
thai	851 Ko	410 Ko	48,2 %

Ordres de grandeur des DELACF

DELACF	Texte	Compressé	Taux
français	22,9 Mo	7,2 Mo	31,4 %
anglais	9,1 Mo	1,7 Mo	18,5 %
grec	11,9 Mo	5,2 Mo	44,0 %

Règles d'application des dictionnaires

- Si trouve un mot qui n'a pas déjà été reconnu par dictionnaire plus prioritaire, alors on l'ajoute au dictionnaire du texte
- Si un mot n'a été reconnu ni comme mot simple, ni comme partie de mot composé, alors il est ajouté à la liste des mots inconnus
- Correspondance minuscules/majuscules

Dictionnaires filtres

3 niveaux de priorités, exprimés par suffixation des noms de fichiers:

dico-.bin	priorité maximum
dico.bin	priorité normale
dico+.bin	priorité minimum

Exemple de filtre

Dans le dictionnaire par défaut (de priorité normale),
`par` peut être un nom (golf)

On peut éliminer cette ambiguïté en utilisant le
dictionnaire prioritaire suivant:

`par , .PREP`

⇒ les entrées de `par` du dictionnaire principal ne
seront plus prises en compte