Lexique et morphologie

22-29 novembre 2021

—Kata Gábor
kata.gabor@inalco.fr—

Langue formelle :
une langue *L*
sur un alphabet/vocabulaire *V*,
est un ensemble de chaînes de symboles produit par concaténation des symboles dans V.
L est ainsi un sous-ensemble de toutes les chaînes que l'on peut produire par concaténation des éléments de *V*.

Grammaire d'une langue formelle :
un modèle capable de reconnaître / générer toutes les chaînes de caractères qui appartiennent à cette langue.

Les automates á états finis et les expressions régulières reconnaissent les mêmes langues : les langues régulières.

Langue régulière :

L est une langue régulière sur l'alphabet V :

- $\emptyset$ (l'ensemble vide/langage vide) est une langue régulière ;
- $\forall a \in V, \{a\}$ (l'ensemble composé d'un élément du vocabulaire) est une langue régulière ;
- Si L_1 et L_2 sont des langues régulières,
  - $L_1 \cup L_2$, **l'union** de L_1 et L_2,
  - $L_1 \cdot L_2$, la **concaténation** de L_1 et L_2, c'est-á-dire $\{xy \mid \mathrm{x} \in L_1, \mathrm{y} \in L_2\}$
  - L_1^* , la **fermeture de Kleene** de L_1, c'est-á-dire la langue obtenue en prenant un nombre quelconque (y compris 0) de chaînes de L_1 et en les concaténant.

Les langues régulières sont fermées pour les opérations de
- complémentation
- inversion des chaînes
- intersection
- différence

Automates et expressions régulières

Signification	**Expression**
séquences de caractères	/chien/, /loup/, /François Hollande/
disjonction	/[aA]/, /[2-5]/, /[a-z]/, / a | b/
négation	/[^a]/
quantification	/a ?/, /b*/, /c+/ [1], /e{3,5}/
parenthèses	/a(ab)+/, /ceci n'(estlétait) pas/
n'importe quel caractère	/./
repères	/^abcdef$/
alias	/\d/, /\s/
caractères spéciaux	/\n/, /\t/

opérations sur des automates	notation
concaténation	A B
optionnalité	A ? = (|A)
fermeture Kleene	A* = (|A|AA|AAA|...)
conjonction	A & B
complément	!A

Construction d'une morphologie computationnelle M

1. créer des automates pour les (classes de) radicaux
2. créer des automates pour les affixes (préfixes, suffixes)
3. combiner les automates de radicaux avec les automates d'affixes (règles morphotactiques : *dans quel ordre les morphèmes se suivent, règles phonologiques et orthographiques*)

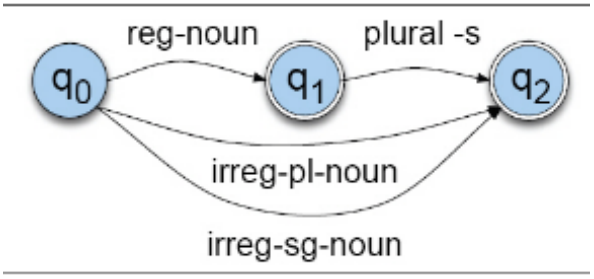

FIGURE 1 – Pluriel régulier et irrégulier des noms

un **Transducteur à états finis (FST)** est constitué de :

— V, un vocabulaire ou alphabet d'entrée
— S, un vocabulaire ou alphabet de sortie
— Q, un nombre fini d'états dont
  — état initial
  — état final (un ou plusieurs)
— f, fonction de transition : donne pour chaque état q $\in$ Q et chaque élément du vocabulaire v $\in$ V, le ou les états que l'on peut atteindre en partant de q et en utilisant v : f(q,v) $\subset$ Q.
— h, fonction de sortie qui associe à un état q et une entrée v $\in$ V une sortie s $\in$ S

Opérations sur les FSTs :

— **inversion** : échanger les étiquettes, l'entrée devient la sortie et vice versa. Un FST analyseur peut être ainsi transformé en FST générateur.
— **union** de deux transducteurs $T_1 \cup T_2$: x [$T_1 \cup T_2$] y si x [T_1] y ou x [T_2] y (le transducteur $T_1 \cup T_2$ traduit x en y si T_1 traduit x en y ou T_2 traduit x en y).
— **composition** de T_1 et T_2 : applique T_1, puis applique T_2 à la sortie de T_1. x [T_1] y puis y [T_2] z : le vocabulaire de sortie de T_1 est le vocabulaire d'entrée de T_2.

Morphologie par FSTs

$$M \subseteq F \times C \tag{1}$$

où F est l'ensemble des formes et C est l'ensemble des descriptions (catégories).

1. le **FST** a un vocabulaire d'entrée et un vocabulaire de sortie
2. **FST** en tant que **traducteur** : de la langue L sur le vocabulaire F (formes de mots) vers la langue L' sur le vocabulaire C (catégories grammaticales et flexionnelles)
3. **FST** définit une langue de paires :

$$F \times C = \{(f, c) | f \in F \ \ et \ \ c \in C\} \tag{2}$$

4. **FST** peut être traduit en **FSA** qui *accepte* une langue :

$$V_{FSA} = \{(f, c) | f \in F \ \ et \ \ c \in C\} \tag{3}$$

   accepte les paires $\{(f, c) \in M \ \}$

Implémentations

— KIMMO, PC-KIMMO : Koskenniemi 1983, Karttunen 1983
— pour la rapidité :
  — XFST : Xerox Finite State Tool
  — SFST : Stuttgart Finite State Tool (morphologie anglaise, allemande, turque, latine disponible)
— pour la facilité et l'interface graphique :
  — Unitex, https ://unitexgramlab.org/fr
  — NooJ, http ://www.nooj-association.org/

Exercices

1. Créez des automates pour les langues suivantes, su le vocabulaire {a,b,c,A,B,C,1,2} :
   (par mot nous entendons des chaînes de caractères séparés par des espaces)

   — l'ensemble de tous les string alphabétiques,
   — l'ensemble de tous les strings alphabétiques en minuscules finissant par b,
   — l'ensemble de tous les strings finissant par deux mots identiques (le dernier mot répété),
   — tous les strings alphabétiques (en 1) directement précédés et suivis par un b,
   — tous les strings alphabétiques en minuscules finissant par b (en 2), directement précédés et suivis par un b (supplémentaire),
   — tous les strings qui commence par une chiffre (p.ex. 1, 2, 222, ...) et qui termine en fin de ligne par un mot.

2. Considérez le tableau de conjugaison allemande du verbe 'singen' (*chanter*) avec les formes en indicatif présent, passé et subjonctif présent.
   (a) Construisez un automate *minimal* qui reconnaît toutes ces formes.
   (b) Pour chaque temps et mode verbal listés, construisez un transducteur individuel qui reconnaît les formes et associe à chaque forme l(es) analyse(s) correspondantes(s).

Personne	Indicatif présent (IPR)	Indicatif passé (IPS)	Subjonctif présent (SP)
S1	singe	sang	sänge
S2	singst	sangst	sängest
S3	singt	sang	sänge
PL1	singen	sangen	sängen
PL2	singt	sangt	sänget
PL3	singen	sangen	sängen

3. Considérez le tableau de déclinaison de l'adjectif hongrois *nagy* (grand).
   (a) Construisez un automate *minimal* qui reconnaît toutes les formes.
   (b) Construisez un transducteur *minimal* qui reconnaît les formes et associe à chaque forme l(es) analyse(s) correspondantes(s).

Cas	Base Singulier	Base Pluriel	Comparatif Singulier	Comparatif Pluriel
Nominatif	nagy	nagyok	nagyobb	nagyobbak
Accusatif	nagyot	nagyokat	nagyobbat	nagyobbakat
Datif	nagynak	nagyoknak	nagyobbnak	nagyobbaknak
Ablatif	nagytól	nagyoktól	nagyobbtól	nagyobbaktól

4. Construisez un modèle morphologique : lexique et automates qui reconnaissent

   — les déclinaisons latines suivantes :

Cas	Singulier	Pluriel
Nominatif	Rosa	Rosae
Accusatif	Rosam	Rosas
Génitif	Rosae	Rosarum
Datif	Rosae	Rosis
Ablatif	Rosa	Rosis

Cas	Singulier	Pluriel
Nominatif	Murus	Muri
Accusatif	Murum	Muros
Génitif	Muri	Murorum
Datif	Muro	Muris
Ablatif	Muro	Muris

   — Transformez les automates en transducteurs capables d'associer une analyse aux formes reconnues.

5. Construisez un FSA pour la dérivatin en -ion/tion en anglais. Transformez-le pour donner l'analyse en sortie : designation = designate.V+tion

   éléments prédictibles :
   — les verbes qui se terminent en [v] reçoivent -ution
   — les verbes qui se terminent en [t], [d] ou [z] reçoivent des suffixes palatalisés en i

designate	designation
unionize	unionization
prosecute	prosecution
resolve	resolution
expedite	expedition
define	definition
absorb	absorption
circumcise	circumcision
decide	decision

éléments imprédictibles :
— -ation ou -ion ? *unionization* vs *prosecution*
— -ation ou -ition ? *proposition* vs *accusation*