Introduction au TAL

21 mai 2010

—Benoît Sagot—

Examen

La durée de l'examen est de 3 heures. Les documents sont autorisés. L'examen est volontairement long, pour ne pas pénaliser outre mesure ceux d'entre vous qui ont assisté à certains cours mais pas à d'autres. Le barème en tiendra compte.

▸ **Exercice 1. Étiquetage morphosyntaxique et reconnaissance d'entités nommées** On rappelle que le terme d'*étiquetage morphosyntaxique*, ou *tagging*, est la tâche consistant à attribuer à chaque token ou à chaque forme d'un texte une *catégorie*, ou *partie du discours*. On peut par exemple définir les catégories `nc` (nom commun), `np` (nom propre), `v` (verbe), `adj` (adjectif), `adv` (adverbe), `prep` (préposition), `det` (déterminant[1]) et `PONCT` (ponctuation).

1. Rappelez la différence entre *tokens* et *formes*. Que sont les *formes composées* et les *amalgames* ?
2. Découpez la phrase ci-dessous en tokens, et attribuez à chacun d'eux une catégorie issue de l'inventaire ci-dessus. On pourra par exemple utiliser une notation de la forme *token*/`cat`. Quel(s) problème(s) rencontrez-vous ? Quelle solution proposeriez-vous ?
   La navette spatiale Atlantis a décollé vendredi du centre spatial Kennedy pour sa 32e et dernière mission, à destination de la Station spatiale internationale (ISS).[2]
3. Transformez les phrases ci-dessous, qui suivent la précédente, en une séquence de formes, et attribuez à chaque forme une catégorie. Certaines formes ne semblent pas rentrer dans l'une des catégories listées ci-dessus. Introduisez de nouvelles catégories lorsque vous le jugez pertinent, en essayant de leur proposer une définition.
   La navette doit livrer à l'ISS un module russe de recherche et d'arrimage. Il s'agit de son dernier vol, même si elle sera ensuite mise en disponibilité comme vaisseau de secours au cas où il faudrait ramener sur Terre un équipage.
4. Rappelez la définition d'une entité nommée. Y en avait-il dans les phrases précédentes ? Si oui, lesquelles ? Découpez la phrase ci-dessous en formes, et attribuez-leur une catégorie, en considérant chaque entité nommée comme une forme unique (éventuellement composée).
   Atlantis s'est élancée du pas de tir de Cap Canaveral à 14h20 locales (18h20 GMT) avec six astronautes à bord.
5. Supposons que l'on dispose d'un *corpus annoté*, c'est-à-dire d'une grande collection de phrases découpées en formes, où l'on a attribué manuellement à chaque forme une catégorie. Quelle technique proposeriez-vous pour exploiter au mieux ce corpus annoté et construire un étiqueteur morphosyntaxique ? On supposera que cet étiqueteur morphosyntaxique n'a pour objectif que d'attribuer à chaque forme une catégorie (le découpage d'un texte quelconque en phrases puis en formes est supposé effectué par un autre module).

[1] Sont des déterminants ce que l'on appelle parfois les « articles » (*les*, *une*) et les « adjectifs déterminatifs » (*ces*, *ton*…).
[2] Reuters | 14.05.10 | 21h50

▸ **Exercice 2. Lexiques et morphologie dérivationnelle** Le développement manuel de ressources lexicales est une tâche fastidieuse. Toute approche permettant d'automatiser tout ou partie du processus est donc la bienvenue — mais nécessite en génréral une validation manuelle.

1. Rappeler brièvement ce qu'est un lexique morphologique, un lexique syntaxique et un lexique sémantique.
2. La *dérivation morphologique* est un processus linguistique permettant de créer un mot à partir d'un mot de base et d'affixes (en français, de préfixes ou de suffixes). Proposez une analyse en base et affixes pour les mots suivants. On indiquera dans chaque cas la catégorie du mot de base et celle du mot dérivé.
   rejouer, *surentraînement*, *déconstruction*, *décomplexer*, *désamiantage*, *chauffagiste*, *droit-de-l'hommisme*, *insoluble*, *réévaluation*
3. Chacun des affixes listés à la question précédente a-t-il un sens similaire d'un exemple à l'autre ? Indiquez quel est cet effet sémantique. Par exemple, le suffixe *-ifier* peut vouloir dire, entre autres, *rendre ADJ* (*complexe* → *complexifier* = *rendre (plus) complexe*) ou *transformer en NC* (*vin* → *vinifier* = *transformer en vin*).
4. Quel est le sens exact de l'expression *insoluble dans l'eau*, et notamment du préfixe que vous avez probablement identifié dans l'adjectif *insoluble* ? Quelle difficulté cela implique-t-il pour l'analyse sémantique d'expressions comme celle-ci ?
5. À l'inverse, comparez les classes flexionnelles[3] de *jouer* et *rejouer* ainsi que leurs cadres de sous-catégorisation[4]. Que pouvez-vous en déduire sur l'impact de la morphologie dérivationnelle pour le développement de lexiques syntaxiques ?

▸ **Exercice 3. Corpus arboré et grammaire d'arbres adjoints**
Rappelons qu'une grammaire d'arbres adjoints (TAG, pour l'anglais *Tree Adjoining Grammar*) est un type de grammaire de réécriture qui permet de combiner des arbres dits *élémentaires* entre eux.

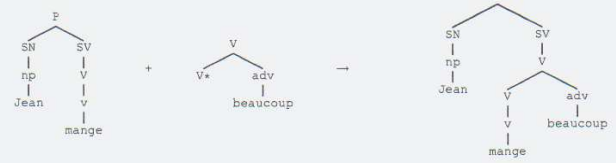
Les arbres élémentaires sont de deux types : les arbres *initiaux* et les arbres *auxiliaires*. Ils se combinent comme suit. Un arbre initial de racine `R` se *substitue* à un *nœud de substitution* d'un autre arbre, c'est-à-dire un nœud feuille non-terminal muni d'un symbole « ↓ ». Un arbre auxiliaire de racine `R`, et qui a nécessairement une feuille `R*`, sépare un nœud étiqueté `R` dans un autre arbre en deux demi-nœuds, le demi-nœud supérieur fusionnant avec la racine de l'arbre auxiliaire et le demi-nœud inférieur fusionnant avec le nœud pied (le nœud étiqueté `R*`).

Voici, pour illustrer, un exemple de chaque opération.

Substitution :

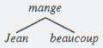
[3] C'est-à-dire, pour un verbe, la façon dont il se conjugue.
[4] Un cadre de sous-catégorisation pour un emploi d'un lemme (ici, d'un verbe) est la liste des fonctions syntaxiques de ses possibles arguments (sujet, objet direct, objet indirect en *de* ou en *à*, etc…) et les réalisations possibles de ces arguments (syntagme nominal, proposition infinitive, syntagme prépositionnel, pronom clitique…)

Adjonction :



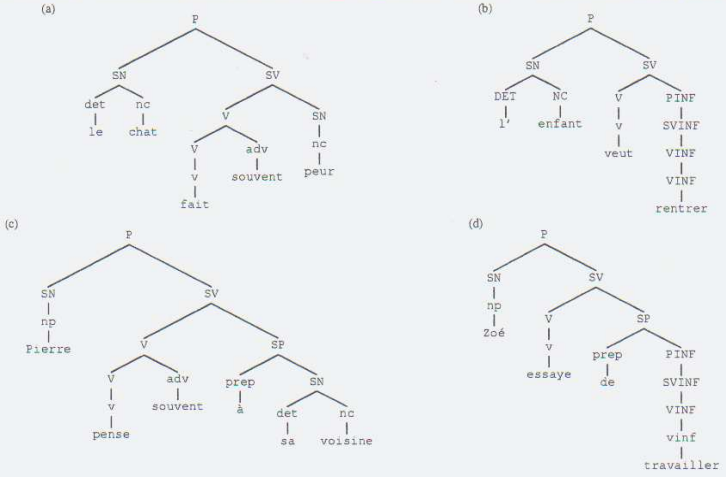
On appelle *ancre* une feuille qui n'est ni un nœud pied ni un nœud de substitution : c'est alors un nœud représentant un mot.

On rappelle enfin que l'ensemble des opérations permettant de construire l'arbre complet d'une phrase à partir des arbres élémentaires peut être représenté par un arbre dit *arbre de dérivation* (l'arbre de la phrase s'appelle *arbre dérivé*). Ainsi, l'arbre de dérivation représentant l'analyse de la phrase *Jean mange beaucoup* au moyen de la séquence des deux opérations ci-dessus est représentée par l'arbre de dérivation suivant :



Cet arbre indique que l'arbre d'ancre `Jean` est substitué dans celui d'ancre `mange`, et que l'arbre d'ancre `beaucoup` y est adjoint. On fait en général en sorte, autant que possible, que l'arbre de dérivation corresponde à un arbre sémantique (le nœud correspondant à un prédicat a pour fils ses arguments et ses modifieurs/circonstants, ce qui est bien le cas ici pour le prédicat *mange*).

Considérons par exemple le corpus arboré suivant, constitué de quatre arbres (et donc d'autant de phrases) :



1. Extraire le lexique correspondant à ce mini-corpus arboré[5]
2. Notre objectif est d'extraire une grammaire d'arbre adjoints de ce corpus. Le lexique extrait à la question précédente permet de s'affranchir de la présence des mots dans les arbres : les terminaux (`det`, `nc`...)

[5]On rappelle qu'un lexique est une liste de couples de la forme (*forme*, *terminal*).

seront les feuilles des arbres de la grammaire. Par ailleurs, nous allons construire une grammaire telle que chaque arbre élémentaire contienne un et un seul nœud ancre.
On suppose que si un nœud N a la même étiquette que son père, c'est qu'il y a eu adjonction d'un arbre auxiliaire. Identifier dans les arbres du mini-corpus les endroits où l'on trouve cette configuration. Combien d'arbres auxiliaires sont-ils en jeu ? Dessinez cet/ces arbres(s) auxiliaire(s), ainsi que l'arbre des phrases correspondantes avant que la/les adjonction(s) ai(en)t eu lieu.

3. Une fois les arbres auxiliaires identifiés et retirés des arbres du mini-corpus, nous allons « découper » ces arbres en arbres initiaux. Proposer un inventaire d'arbres initiaux couvrant exactement le mini-corpus arboré, en indiquant le nombre de fois que chacun d'eux est utilisé dans le mini-corpus. On prendra garde à faire en sorte que lorsque l'on reconstruit l'arbre complet d'une phrase, l'arbre de dérivation obtenu soit satisfaisant (le nœud pour l'arbre ancrant le verbe principal doit par exemple être la racine de l'arbre de dérivation).
4. Normalisez les comptes ainsi obtenus sur les arbres initiaux pour en faire des probabilités de réécriture : la somme des probabilités des arbres initiaux de même racine doit être égale à 1.
5. Que proposez-vous comme probabilités pour le(s) arbre(s) auxiliaire(s) ?
6. Calculez les probabilités globales de chacun des quatre arbres du mini-corpus.
7. Soit la phrase *Pierre commence à dîner*. Complétez le lexique pour qu'il couvre tous les mots de cette phrase (on n'oubliera pas que *dîner* est à la fois un nom commun et un infinitif).
8. Combien d'analyses cette phrase a-t-elle selon la grammaire d'arbres adjoints que vous avez extraite précédemment du mini-corpus arboré ? Quelles sont les probabilités associées à ces analyses ? L'analyse la plus probable est-elle la bonne ?
9. Quelle est votre impression générale sur cette approche ? Quelles idées auriez-vous pour la rendre plus sophistiquée ? Quels avantages et/ou inconvénients lui trouvez-vous par rapport aux PCFG ?