2 Etiquetage morpho-syntaxique

(Garside et al. 1997, chapitre 2)

Comprend deux questions:

1. Quel ensemble d'étiquettes (*tagset*) faut-il choisir?
2. Comment lier les tokens à leurs étiquettes?

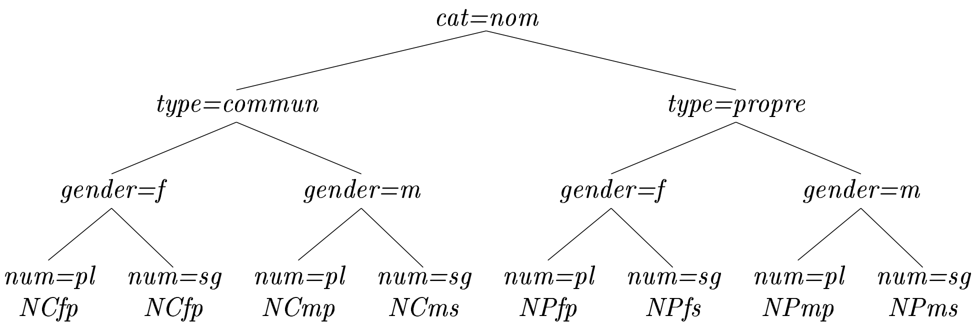
2.1 Tagsets: Choix des catégories morpho-syntaxiques

Distinction entre la catégorie morpho-syntaxique et l'étiquette utilisé pour cette catégorie. Par exemple, `P` ou `Prep` peuvent être des étiquettes pour la catégorie "préposition".

Tagsets logiques: Les relations entre catégories morpho-syntaxiques dans le tagsets peuvent être representées dans un arbre hiérarchique.

Un tagset logique peut être formulé en termes d'une hiérarchie d'attributs et valeurs.

Exemple 4 *(partie du tagset du corpus Le Monde, TALaNa):*



En principe chaque tagset peut être représenté comme tagset logique. La question intéressante est si le tagset a été developpé comme tagset logique et si cela se voit à partir des étiquettes choisies (comme dans l'exemple).

Pour la conception d'un tagset, il y a différents aspects à considérer:

(a) les distinctions linguistiques qu'on veut faire

(b) des traits distributionnels des catégories (pour un tagging automatique il est mieux de choisir des catégories qu'on peut déterminer à partir des catégories avoisinantes)

Exemple 5 *On aimerait du point de vue linguistique distinguer entre des verbes en subjonctif présent et des verbes en indicatif présent. Mais comme les catégories avoisinantes sont souvent les mêmes, un étiqueteur automatique va probablement produire des erreurs en faisant cette distinction. Il vaut peut-être mieux ne pas distinguer entre les deux formes pour obtenir une plus grande précision de tagging.*

Recommandations EAGLES pour l'annotation morphosyntaxique (voir lien sur page web du cours):

Tagset représenté comme tagset logique avec des attributs/valeurs. Distinction entre

1. des attributs/valeurs obligatoires: seulement les catégories principales:
   N (nom), V (verbe), AJ (adjectif), PD (pronom/déterminant), AT (article), AV (adverbe), AP (adposition), C (conjonction), NU (nombre), I (interjection), U (unique), R (reste), PU (ponctuation).

2. des attributs/valeurs recommandés pour chaqu'une des catégories principales.
   Idée: si un de ces attributs occure dans le langage du texte à annoter, il devrait être encodé dans le tagset.
3. des attributs optionnels pour chaqu'une des catégories principales: des attributs spécifics pour certains buts de recherches ou spécifics pour certains langages.

Exemple 6 *Pour la catégoie N, EAGLES recommande les attributs/valeurs suivants:*

1. *"type" avec valeurs common ou proper*
2. *"gender" avec valeurs masculine, feminie ou neuter*
3. *"number" avec valeurs singular ou plural*
4. *"case" avec valeurs nominative, genitive, dative, accusative ou vocative*

Comme il n'y que féminin et masculin en français, et comme il n'y a pas de cas pour les noms, les tags distingués pour les noms dans le corpus Le Monds (TALaNa) sont conformes aux recommandations EAGLES.

Tagset du Brown Corpus: 87 catégories

Tagset du Penn Treebank: 48 catégories

2.2 POS Tagging

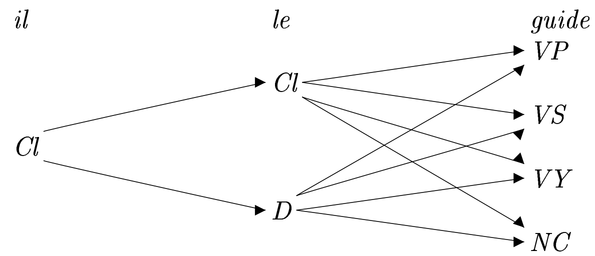
Le procès d'attribuer des catégories morpho-syntaxiques (parties du discours) aux mots (tokens) d'un texte s'appelle tagging ou POS tagging.

Le tagging peut être fait presque automatiquement. Les étiqueteurs aujourd'hui ont normalement une précision de plus que 95%. Après le tagging automatique il faut une correction manuelle pour augmenter la précision.

Les catégories possibles pour un mot sont extraites d'un dictionnaire.

Problème: Les mots/tokens sont souvent ambigus à l'ègard de leur catégorie.

Exemple 7 *Ambiguités dans "il le guide" (étiquettes comme dans le corpus Le Monde TALaNa)*



(VP = verbe présent indicatif, VS = verbe présent subjonctif, VY = verbe impératif)

Pour cette phrase, il y a $1 \times 2 \times 4 = 8$ *suites de catégories possibles. (8 chemins traversants le graphe orienté.)*

Désambiguisation à partir du contexte locale. Deux approches:

(a) par règles ou
(b) probabiliste

2.2.1 Désambiguïsation par règles

Des grammaires locales qui essaient de désambiguiser à partir de n (1 ou 2 normalement) catégories precédentes et suivantes.

Règles négatives: règles qui excluent des suites de catégories. P.e. une règle excluante une suite CL NC ou une règle excluant une suite D V. Ces deux règles excluent déjà la moitié des chemins dans le graphe dans ex. 7.

Règles positives: règles qui déterminent une catégorie. P.e. une règle qui dit qu'un mot qui peutavoir une catégorie V ou N et qui est suivi de "-t-il" a la catégorie V.

Des suites désambiguisées s'appellent "ilôts de confiance".

Problèmes:

- on peut appliquer les règles seulement si le contexte n'est pas ambigue
- il faut créer les règles manuellement
- les règles dépendent du tagset
- on ne peut décrire que des phenomènes locaux

2.2.2 Désambiguïsation probabiliste

But: déterminer le chemin le plus probable dans le graphe orienté des étiquettes possibles

Hidden Markov Model.

La probabilité d'une suite de catégories $c_1, \ldots, c_n$ pour une suite de mots $m_1, \ldots, m_n$ est calculée en fonction de

1. la probabilité $P(m_i|c_i)$ de la catégorie c_i pour le mot m_i et
2. la probabilité $P(c_i|c_{i-1})$ ou $P(c_i|c_{i-1}, c_{i-2})$ de la catégorie c_i à partir de c_{i-1} (bigrame) ou de c_{i-1} et c_{i-2} (trigrame).

$$\begin{tikzcd} m_1 & m_2 & \cdots & m_{n-1} & m_n \\ c_1 \arrow[u, "q_1"] \arrow[r, "p_{1,2}"'] & c_2 \arrow[u, "q_2"] & \cdots & c_2 \arrow[u, "q_{n-1}"] \arrow[r, "p_{n-1,n}"'] & c_n \arrow[u, "q_n"] \end{tikzcd} \qquad q_i = P(m_i|c_i) \text{ et } p_{i-1,i} = P(c_i|c_{i-1})$$

Les probabilités sont calculées en utilisant un corpus d'apprentissage (un corpus étiqueté).

Probabilité d'une catégorie c pour un mot m:

$$P(m|c) = \frac{freq(m|c)}{freq(m)}$$

Probabilité que la catégorie c_i suit la catégorie c_{i-1}:

$$P(c_i|c_{i-1}) = \frac{freq(c_{i-1}c_i)}{freq(c_{i-1})}$$

Problèmes: la probabilité d'une suite de catégories qui ne figure pas dans le corpus d'apprentissage est 0. Pour empêcher cela, on attribue dans ces cas une probabilité minimale.

Avantages des approches probabilistes:

- robuste
- apprentissage automatique
- peut être appliqué à d'autres corpus
- en general meilleure précision qu'avec désambiguïsation par règles

Problèmes:

- le corpus d'apprentissage doit être relativement grand pour fournir des probabilités fiables.
- ne décrit que des dépendances locales