- Vous avez le droit à vos notes de cours, livres, etc.
- Vous pouvez visualiser vos notes de cours sur votre ordinateur portable. Cependant, vous devez désactiver la connexion au réseau et ne faire usage d'aucune autre application que votre lecteur pdf. Votre téléphone ne peut être utilisé que pour vous donner l'heure.
- Répondez directement sur le carnet de réponse; les questions appellent le plus souvent à des réponses courtes et précises.

---

1. **Modèle de langue**
   (a) Lesquels de ces modèles consultent **toujours** le modèle d'ordre inférieur:
      (1) Katz
      (2) Jelinek-Mercer
   (b) Indiquez les modèles qui donnent une probabilité fixe pour tous les mots inconnus:
      (1) Add-one
      (2) Jelinek Mercer
      (3) Maximum de vraisemblance
      (4) Good Turing
   (c) Considérez les 2 modèles probabilistes discrets suivants définis sur le même domaine:

	a	b	c	d
p_1	1/2	1/6	1/6	1/6
p_2	1/4	1/4	1/6	1/3

      (1) Lequel de p_1 et p_2 a la plus grande entropie ?
      (2) Exprimez l'entropie croisée $H(p_1, p_2)$ en fonction de ces probabilités (je ne vous demande pas de la calculer)
      (3) $H(p_2, p_1)$ est-elle plus grande ou plus petite que l'entropie de p_2 ?
   (d) Quel est le rapport entre un modèle de langue et le concept d'entropie ?

2. **Analogies**

   On s'intéresse ici à la définition d'analogie formelle de Stroppa & Yvon.
   (a) Identifiez les analogies de formes parmi les relations suivantes :
      (1) aaaa : bb :: abab : baba
      (2) postier: osti: pier : i
      (3) aba : baa : bab : aba
      (4) run : ran :: aratun : arutun
      (5) run : ran :: arutun : aratun
   (b) Résoudre les équations analogiques formelles suivantes:
      (1) [aaaa: ? :: abab : baba]
      (2) [abab: aaaa :: bbbb : ?]

3. **Étiquetage**

   (a) Indiquez la séquence d'étiquettes *IOB* qui permet de représenter cette phrase de neuf mots (les mots sont délimités par un espace):

   (Le grand homme)$_{np}$ boit (de l' eau forte)$_{NP}$.

   (b) Comment appelle t-on le résultat d'un tel étiquetage ?

   (c) Quel est le jeu d'étiquettes utilisé dans les *benchmarks* de CONLL pour la reconnaissance d'entités nommées ?

4. **Grammaire**

   Considérez la grammaire G dont S est l'axiome:

$$\begin{aligned} &\mathrm{S} \rightarrow \mathrm{AS} \mid \epsilon \\ &\mathrm{A} \rightarrow 0\mathrm{A}1 \mid \mathrm{A}1 \mid 01 \end{aligned}$$

   (a) G est-elle récursive à gauche ? Justifiez.

   (b) G est-elle une grammaire régulière ? Justifiez.

   (c) G est-elle LL1? Justifiez.

   (d) Quel est le langage décrit par le non terminal A de la grammaire G ?

   (e) Quel est le langage décrit par G ?

   (f) Peut-on décrire un language de type n à l'aide d'une grammaire de type m où $n > m$? Justifiez.

   (g) Peut-on décrire un language de type n à l'aide d'une grammaire de type m où $n < m$? Justifiez.

   (h) Construisez la table d'analyse de l'algorithme d'Earley pour l'analyse de la phrase: *00111011*. Chaque item considéré par l'algorithme doit être indiqué.

   (i) Écrire une grammaire qui décrit le language de toutes les chaînes sur l'alphabet $\{a, b, c\}$ ne contenant pas la séquence *abc*. `ab`, `abac` sont des exemples de chaînes appartenant à ce langage.

   (j) Le language décrit en i) est-il selon vous régulier ? Justifiez.

   (k) Expliquez en une phrase en quoi consiste CYK.

5. **Modèle de Markov**

   Considérez M un modèle de Markov à 3 états qui génère des symboles sur l'alphabet $\{a, b, c, d\}$. La probabilité de commencer par l'un de ces états est la même pour tous les états ($\pi_i = 1/3, \forall\, i\, [1, 3]$).

   (a) En admettant que chaque état puisse émettre chaque symbole, combien de séquences d'états peuvent expliquer une observation de 4 symboles? Justifiez.

   (b) Que calcule l'algorithme de Viterbi ?

(c) Combien d'additions l'algorithme de Viterbi effectuera t-il pour la séquence *abbbcd*? Justifiez.

(d) Vous souhaitez ajuster les probabilités de M de manière à en faire un modèle de langue qui reconnaît toutes les chaînes sur l'alphabet $\{a, b, c, d\}$ tout en favorisant les chaînes qui commencent par un a et qui contiennent les symboles en ordre croissant (*abd* serait par exemple préférable à *adb*). Indiquez les paramètres de votre modèle.

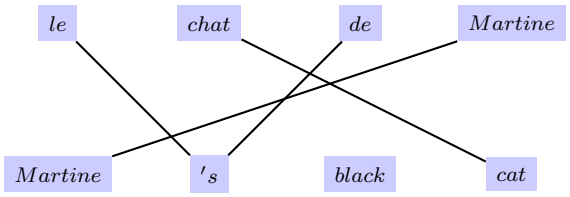
6. **Modèles IBM**

(a) Quelle est la contrainte d'alignement introduite dans les modèles de traduction IBM?

(b) Par quelle structure de donnée est représenté un alignement dans un tel modèle ?

(c) Pourquoi cette contrainte a t-elle été proposée? Comment sont gérés les mots non alignés dans ce type d'alignement?

(d) L'alignement suivant a t-il été obtenu avec un modèle IBM $p(e|f)$ ou $p(f|e)$, e et f étant des mots respectivement en anglais et en français ?



(e) Exprimez la probabilité de cet alignement, selon le modèle IBM2 (exprimer $\neq$ calculer).

(f) Vous disposez de deux corpus de textes, l'un en anglais, l'autre en français. Ces deux corpus sont parallèles. Indiquez une façon d'obtenir à l'aide des modèles IBM des synonymes des mots de la langue anglaise.

7. **Distance d'édition**

On supposera dans cette question les coûts d'insertion, de suppression et de substitution unitaires.

(a) La distance d'édition entre deux chaines peut-elle être plus grande que la plus grande des deux chaînes? Justifiez.

(b) Représentez la table d'édition pour les châines ABCD et ACBD. Combien existe-t-il d'alignements de coût minimum ? Indiquez les.

(c) Quelle est la distance de Hamming entre ces deux chaines?

(d) Vous souhaitez introduire une distance d'édition qui reconnait les opérations de permutation locales (deux symboles adjacents). On peut par exemple passer de `corss` à `cross` à l'aide d'une telle opération. Quelle condition sur le coût de cette opération devez vous observer pour que cette opération soit utilise ?

(e) Écrire un algorithme capable de calculer une telle distance (avec l'opération de la question précédente).