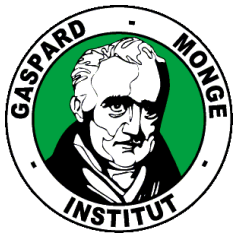
Master 1 Informatique – Université Marne-la-Vallée (IGM)
20/03/2014 – Cours 6
Ingéniérie Linguistique

Traduction automatique



UNIVERSITÉ
PARIS-EST
MARNE-LA-VALLÉE

Philippe Gambette

Annonces

- Annales d'examen
- Pour aller plus loin :
  - ATALA (depuis 1959), liste LN http://www.atala.org/

    Association pour le traitement automatique des langues
  - AFTAL : http://tal.univ-paris3.fr/aftal/

    Anciens des Formations TAL

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Sources

• Cours de Matthieu Constant, *Ingéniérie Informatique 1*

http://igm.univ-mlv.fr/ens/Master/M1/2010-2011/IngenierieLinguistique1/cours.php

• Daniel Jurafsky and James H. Martin, 2007, *An introduction to natural language processing, computational linguistics and speech recognition*

• Christopher D. Manning and Heinrich Schütze, 1999, *Foundations of Statistical Natural Language Processing*, Massachussetts Institute of Technology

• Ruslan Mitkov, 2003, *The Oxford Handbook of Computational Linguistics*, Oxford University Press

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Motivations

Objectif

Traduire un texte d'une langue source dans une langue cible automatiquement.

Exemple

- Anglais : The poor don't have any money.
- Français : Les pauvres sont démunis.

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Problèmes

Ambiguïté de la langue source

- Ambiguïté grammaticale
    - Exemple 1 : *light* est soit un nom (*lumière*), soit un verbe (*allumer*), soit un adjectif (*clair* ou *léger*)
    - Exemple 2 : *face* est soit un nom (*visage*), soit un verbe (*affronter*)
- Ambiguïté syntaxique
- Ambiguïté sémantique
    - Exemple 1 : *voler* → *fly* ou *steal*
    - Exemple 2 : *bank* → *banque* ou *berge*
- Ambiguïté de référence des pronoms
- …

Problèmes

Structures

The student **is likely to** work

= **Il est probable que** l'étudiant travaillera

Anaphores

The soldiers killed the women. **They** were buried the next day.

= Les soldats ont tué les femmes. **Elles** (*Ils) furent enterrées le jour suivant.

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Trois approches

Approche interlangue

1. Représentation syntactico-sémantique interlangue du texte source
2. Génération du texte cible en partant de la représentation

Approche par transfert

1. Analyse lexicale et syntaxique du texte source
2. Transfert des structures et des traductions lexicales en langue cible

Approche directe

1. Traduction mot à mot du texte source vers le texte cible
2. Modification de l'ordre des mots traduits dans le texte cible

Approche par transfert

Transfert lexical

• Exemple : *car.Noun* ↔ *voiture.Noun*

• Besoins : dictionnaires bilingues

Transfert de structures

• Exemple : *the Adj Noun* ↔ *le Noun Adj*

• Besoins : analyseur syntaxique et formalisme de transfert
(ex. transducteurs d'arbres)

Transfert lexico-syntaxique

• Exemple : NP consist of NP ↔ NP consister en NP

• Besoins : idem que pour les transferts de structures + lexiques syntaxiques

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Traduction automatique statistique

Objectif

Étant donné une phrase *F* dans une langue source (ex. français, espagnol), le but est de trouver une phrase $\hat{E}$ en langue cible (ex. anglais) qui maximise la probabilité conditionnelle d'avoir une phrase *E* en langue cible.

Formule

$$\hat{E} = \operatorname{argmax}_E P(E|F)$$

$$= \operatorname{argmax}_E \frac{P(F|E).P(E)}{P(F)} \text{ (Formule de Bayes)}$$

$$= \operatorname{argmax}_E P(F|E).P(E)$$

Traduction automatique statistique

Un modèle de traduction automatique statistique nécessite trois composants

- Un **modèle de langage** qui calcule P(*E*)
- Un **modèle de traduction** qui calcule P(*F*|*E*)
- Un **décodeur** qui prend une phrase *F* et produit la phrase la plus probable *E*.

Modèle de traduction

Modèle basé sur les mots

- Calculer des probabilités de traductions des mots
- Calculer des probabilités de déplacement des mots en langue cible

Modèle basé sur les séquences de mots

- idem mais pour les séquences de mots

Modèle fondé sur la syntaxe

- Mise en parallèle de structures d'arbres syntaxiques
- Modèle mathématique : grammaire transductive

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Modèle basé sur les séquences de mots

Séquence de mots = unité fondamentale de traduction

On utilise une table de traduction de séquences.

Par exemple, trad["green witch"] = [("grüne Hexe",0.86),...]

Génération de la traduction

Pour chaque phrase $E = e_1 e_2 \ldots e_n$, simultanément,

- on regroupe les mots en séquences ($E = \bar{e}_1 \bar{e}_2 \ldots \bar{e}_k$) (plusieurs possibilités !)
- on traduit chaque séquence $\bar{e}_i$ en une séquence $\bar{f}_j$
- on change, si besoin, l'ordre des mots dans *F*

Modèle basé sur les séquences de mots

Exemple :

- anglais - allemand

The green witch | is | at home | this week

Diese Woche | ist | die grüne Hexe | zu Hause

Composants du modèle de traduction

Probabilité de traduction $\phi(\bar{f}|\bar{e})$

- Probabilité de générer $\bar{f}$ à partir de $\bar{e}$
- Exemple : ϕ(“green witch”|“grüne Hexe”) = 0.86

Probabilité de distorsion $d(n)$

- Probabilité que les traductions de deux séquences consécutives de mots soient distantes de n dans la phrase générée
- Distance calculée en nombre de mots

Apprentissage

Idéalement

• Utilisation d'un corpus bilingue aligné par séquences de mots

• L'estimation des probabilités est alors directe :

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}|\bar{e})}{\sum_{\bar{g}} \text{count}(\bar{g},\bar{e})}$$

Dans la réalité

• Un tel corpus n'existe pas ou est trop petit !

• Extraction automatique des alignements par séquences à partir de l'alignement automatique par mots

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Alignement mot à mot

Corpus parallèle

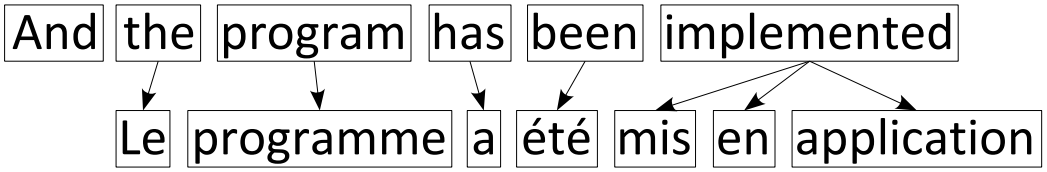
• Un ensemble de couples de textes tel que, pour chaque couple, un des textes est la traduction de l'autre.

• Exemples : Europarl, Hansard

Alignement au mot

Mise en correspondance des mots en langue source avec les mots en langue cible dans un ensemble de phrases parallèles

Exemple anglais → français



Obtention d'un corpus parallèle aligné

1. Découper le corpus en phrases
2. Aligner les textes par phrases (alignements N à M possibles)
3. Pour les couples de phrases alignées, alignement au mot.

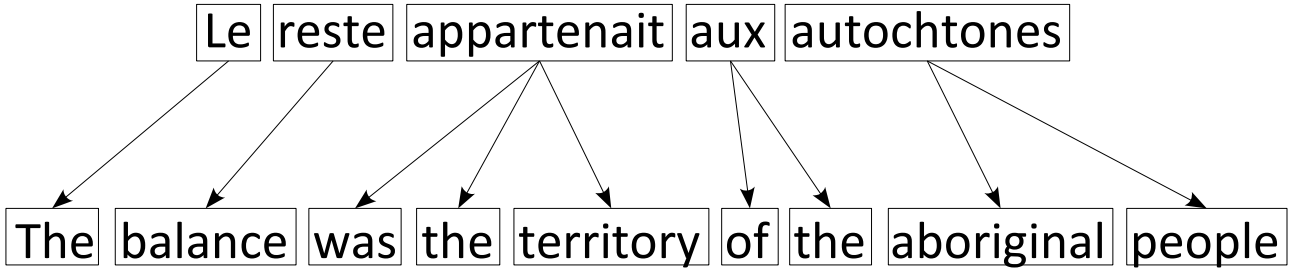
Alignement N à 1

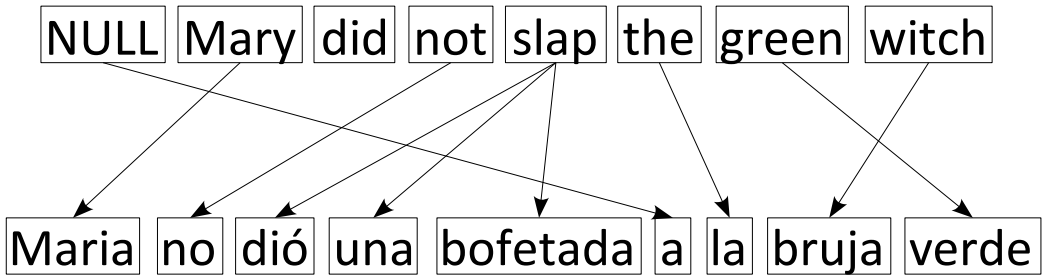
Chaque mot dans la phrase cible ne doit être aligné qu'avec un seul mot dans la phrase source.

Mots vides

Un mot de la phrase cible peut ne pas avoir de correspondant dans la phrase source. On suppose l'existence d'un mot vide (NULL) en langue source aligné avec de tels mots.

Alignement mot à mot





Alignement mot à mot

Un alignement *A* est une suite

À chaque position *j* dans la phrase cible, on associe une position A_j dans la phrase source.

Exemples

- français → anglais : *A* = 1,2,3,3,3,4,4,5,5
- anglais → espagnol : *A* = 1,3,4,4,4,0,5,7,6

Plan

- Introduction
- Problèmes
- Différentes approches
- Traduction automatique statistique
- Modèle de traduction basé sur les séquences
- Alignement mot à mot
- Modèle par heuristiques

Modèle par heuristiques : principe

Fonction de similarité

- Utilisation d'une fonction de similarité entre les mots source et cible
- Basée sur la cooccurrence de ces deux mots dans des phrases alignées

Évaluation

- Résultats moins bons que des modèles probabilistes (ex. logiciel Giza++)
- Temps d'apprentissage tout à fait raisonnables !

Modèle par heuristiques : apprentissage

Notations

• *e* et *f* des mots respectivement en langue source et cible

• *C*(*e*,*f*), le nombre de fois que *e* et *f* apparaissent ensemble dans une paire de phrases alignées

• *C*(*e*), le nombre d'occurrences de *e* dans le corpus en langue source

• *C*(*f*), le nombre d'occurrences de *f* dans le corpus en langue cible

Coefficient de Dice

$$\text{dice(e,f)} = \frac{2\ C(e,f)}{C(e)+C(f)}$$

Modèle par heuristiques : décodage

Matrice de similarité

• Construction d'une matrice de similarité *M* pour chaque paire de phrases à aligner

• À chaque paire de mots (e_i,f_j) des deux phrases alignées, $M(i,j)$ = dice(e_i,f_j)

Différentes heuristiques

• Recherche du maximum :

Pour chaque position *j* dans la phrase cible, a_j correspondra à la position *i* dans la phrase source, qui maximise $M(i,j)$.

• Competitive linking algorithm :

1. Aligner la paire de mots (i,j) qui maximise $M(i,j)$
2. Supprimer la ligne *i* et la colonne *j* de la matrice *M*
3. S'il reste des mots à aligner, aller à 1.

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	0.550	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	0.351	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	0.300

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	0.351	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	0.300

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	0.351	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	0.300

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	0.300

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	0.300

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	0.284	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	0.217	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	**0.284**	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	**0.217**	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	**0.284**	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	0.156	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	**0.217**	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	**0.284**	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	**0.156**	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	**0.217**	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	**0.284**	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	0.126	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	0.128	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	**0.156**	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	0.076	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**

Modèle par heuristiques : décodage

	Le	procès-verbal	de	la	séance	d	hier	a	été	distribué
The	0.193	0.002	0.228	**0.217**	0.006	0.193	0.005	0.143	0.088	0.001
Minutes	0.003	**0.550**	0.002	0.002	0.309	0.002	0.433	0.003	0.003	0.000
of	0.104	0.002	**0.284**	0.228	0.007	0.276	0.006	0.231	0.109	0.001
yesterday	0.006	0.016	0.005	0.006	0.010	0.005	**0.351**	0.006	0.006	0.007
s	0.069	0.003	0.141	0.137	0.006	**0.126**	0.007	0.106	0.070	0.001
sitting	0.008	0.024	0.006	0.006	**0.128**	0.006	0.023	0.007	0.106	0.000
have	0.085	0.003	0.258	0.147	0.007	0.117	0.006	**0.156**	0.088	0.001
been	0.062	0.002	0.121	0.119	0.006	0.113	0.006	0.099	**0.076**	0.000
distributed	0.001	0.012	0.001	0.001	0.000	0.001	0.005	0.001	0.001	**0.300**