CC BY NC SA

# De l'écrit à l'Information: mot, forme, token, fragment

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

# Quelques sources d'inspiration

- Bruno Guillaume : https://grew.fr/download/LIFT_2024_guillaum.pdf

# Retour sur le TD précédent

# Exercice : transcription

Écouter et transcrire

# Retour sur l'exercice

- la transcription est une annotation (besoin de conventions)

# Retour sur l'exercice

- la transcription est une annotation (besoin de conventions)
- il peut y avoir des bruits autre que la parole (applaudissements, rire, etc)

# Retour sur l'exercice

- la transcription est une annotation (besoin de conventions)
- il peut y avoir des bruits autre que la parole (applaudissements, rire, etc)
- plusieurs personnes peuvent parler en même temps

# Retour sur l'exercice

- la transcription est une annotation (besoin de conventions)
- il peut y avoir des bruits autre que la parole (applaudissements, rire, etc)
- plusieurs personnes peuvent parler en même temps
- il n'y a pas de phrase dans la parole

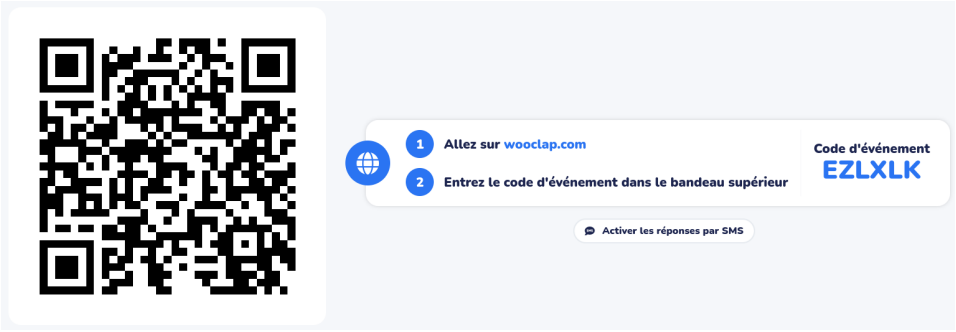
# Segmentation en phrases (texte brut)

- Se termine par une ponctuation de fin de phrase (. ; ? !)
- La phrase suivante commence par une espace ou un retour à la ligne suivi d'une majuscule
- Les . . .forment une ponctuation unique

# Segmentation en phrases : plus loin

- en allemand ?
- en thaï ?

# Wooclap


1 Allez sur wooclap.com
2 Entrez le code d'événement dans le bandeau supérieur
Code d'événement
EZLXLK
Activer les réponses par SMS

# On se réveille (wooclap)

**Exercice**

Combien comptez-vous de mots dans la phrase : "Je suis tout à fait d'accord !" ?

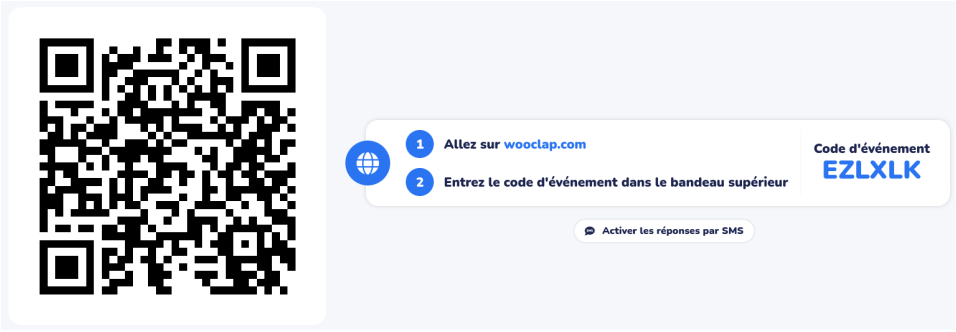
# Unité lexicale ou lexème

*[L'unité est] une tranche de sonorités qui est, à l'exclusion de ce qui précède et de ce qui suit, le signifiant d'un certain concept. [de Saussure, 1916, p. 146]*

# Unité typographique, mot typographique ou token

*[...] un token est une séquence contiguë de caractères délimités de part et d'autre par un séparateur typographique ou par un signe de ponctuation [Sagot, 2018, p. 35]*

# Wooclap


1 Allez sur wooclap.com
2 Entrez le code d'événement dans le bandeau supérieur
Code d'événement
EZLXLK
Activer les réponses par SMS

# On se réveille (wooclap)

**Exercice**

Combien comptez-vous de tokens dans la phrase : "Je suis tout à fait d'accord !" ?

# Correction

"Je suis tout à fait d'accord !" :

- hors ponctuation (" !") :
  - 7 tokens
  - 5 unités lexicales, dont 1 unité polylexicale : "tout à fait"

# Unités polylexicales

**Définition**

Une suite de tokens dont le sens n'est pas compositionnel (on ne peut pas le déduire du sens des tokens qui la compose).
Exemples : "au fur et à mesure", "casser sa pipe"

# Les mots en chinois (simplifié)

de Hee-Soo Choi, avec son accord

Un lexème, un caractère :

- 书 : shū, livre (peut aussi signifier "écrire")

Avec des combinaisons :

# Les mots en chinois (simplifié)

de Hee-Soo Choi, avec son accord

Un lexème, un caractère :

- 书 : shū, livre (peut aussi signifier "écrire")
- 店 : diàn, magasin

Avec des combinaisons :

# Les mots en chinois (simplifié)

de Hee-Soo Choi, avec son accord

Un lexème, un caractère :

- 书 : shū, livre (peut aussi signifier "écrire")
- 店 : diàn, magasin
- 虫 : chóng, insecte/vers

Avec des combinaisons :

# Les mots en chinois (simplifié)

de Hee-Soo Choi, avec son accord

Un lexème, un caractère :

- 书 : shū, livre (peut aussi signifier "écrire")
- 店 : diàn, magasin
- 虫 : chóng, insecte/vers

Avec des combinaisons :

- 书店 (shūdiàn) : librairie

# Les mots en chinois (simplifié)

de Hee-Soo Choi, avec son accord

Un lexème, un caractère :

- 书 : shū, livre (peut aussi signifier "écrire")
- 店 : diàn, magasin
- 虫 : chóng, insecte/vers

Avec des combinaisons :

- 书店 (shūdiàn) : librairie
- 书虫 (shūchóng) : parasite du livre

# Amalgames

Passez au singulier :

1. Ida achète des livres.
2. Ida parle des livres.

# Amalgames

Passez au singulier :

1. Ida achète des livres. → un livre
2. Ida parle des livres.

# Amalgames

Passez au singulier :

1. Ida achète des livres. → un livre
2. Ida parle des livres. → d'un livre

# Amalgames

Passez au singulier :

1. Ida achète des livres. → un livre
2. Ida parle des livres. → d'un livre

Dans le cas 2, "des" est un amalgame (de+le)

# Fragments


https://tiktokenizer.vercel.app

# Fragments (encore)


https://tiktokenizer.vercel.app

# Parenthèse sur la tokenization dans les grands modèles de langues

Ici sur l'interface de test du tokenizer d'OpenAI (H. de Mazancourt)


https://www.linkedin.com/in/mazancourt/

# Fragments : la tokenization est biaisée

*"there's a correlation between worse performance, higher prices, and less advantaged/rich communities." [Ahia et al., 2023]*

# Flexion : ajout d'un affixe qui ne crée pas un nouveau lexème

Pour marquer :

- le genre : instituteur - institutrice

→ le français est une langue flexionnelle

# Flexion : ajout d'un affixe qui ne crée pas un nouveau lexème

Pour marquer :

- le genre : instituteur - institutrice
- le nombre : cheval - chevaux

→ le français est une langue flexionnelle

# Flexion : ajout d'un affixe qui ne crée pas un nouveau lexème

Pour marquer :

- le genre : instituteur - institutrice
- le nombre : cheval - chevaux
- le temps ou la personne (pour les verbes) : mangera - mangeaient

→ le français est une langue flexionnelle

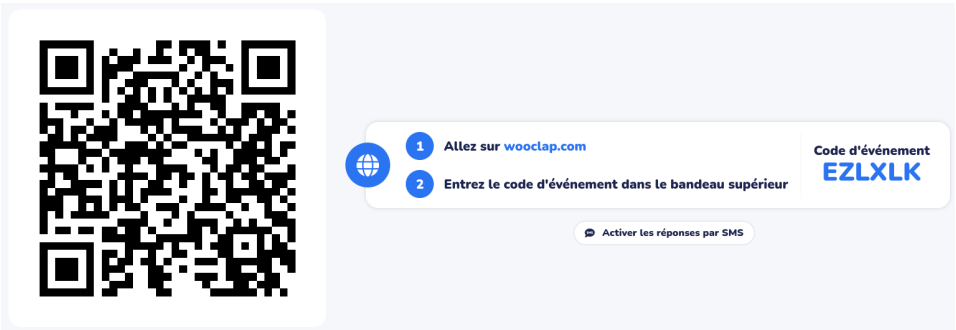
# Flexion : ajout d'un affixe qui ne crée pas un nouveau lexème

Pour marquer :

- le genre : instituteur - institutrice
- le nombre : cheval - chevaux
- le temps ou la personne (pour les verbes) : mangera - mangeaient
- la fonction syntaxique (langues à cas) : lupus (latin, sujet)

→ le français est une langue flexionnelle

# Wooclap


1 Allez sur wooclap.com
2 Entrez le code d'événement dans le bandeau supérieur
Code d'événement
EZLXLK
Activer les réponses par SMS

# On se réveille (wooclap)

**Exercice**

Citez d'autres langues flexionnelles

# Dérivation : ajout d'un affixe qui crée un nouveau lexème

Avec :

- des préfixes : impossible, retirer, extraordinaire

# Dérivation : ajout d'un affixe qui crée un nouveau lexème

Avec :

- des préfixes : impossible, retirer, extraordinaire
- des suffixes : mangeable, tablette

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom :
- Verbe+Verbe :
- Nom+Nom :
- Adj+Nom :
- Nom+Adj :
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe :
- Nom+Nom :
- Adj+Nom :
- Nom+Adj :
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom :
- Adj+Nom :
- Nom+Adj :
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom : loup-garou, chef-lieu
- Adj+Nom :
- Nom+Adj :
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom : loup-garou, chef-lieu
- Adj+Nom : grand-père
- Nom+Adj :
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom : loup-garou, chef-lieu
- Adj+Nom : grand-père
- Nom+Adj : coffre-fort
- Nom+Prép+Nom :
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

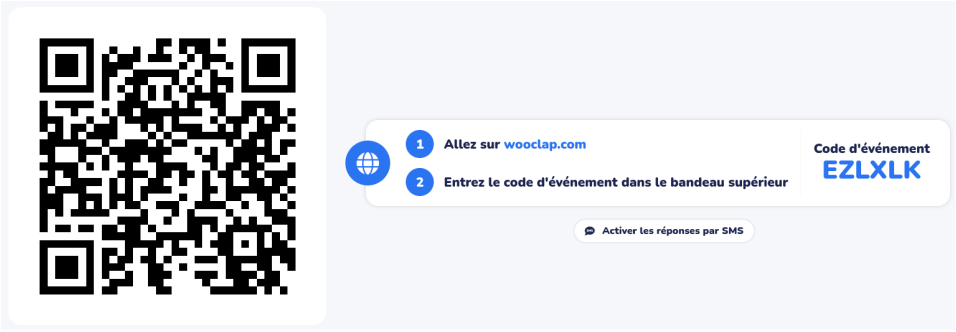
- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom : loup-garou, chef-lieu
- Adj+Nom : grand-père
- Nom+Adj : coffre-fort
- Nom+Prép+Nom : pot-au-feu
- Adj+Adj :

# Composition : création d'un nouveau lexème à partir d'au moins deux

On peut composer des lexèmes ayant des catégories grammaticales différentes :

- Verbe+Nom : porte-manteau, porte-mine
- Verbe+Verbe : savoir-vivre
- Nom+Nom : loup-garou, chef-lieu
- Adj+Nom : grand-père
- Nom+Adj : coffre-fort
- Nom+Prép+Nom : pot-au-feu
- Adj+Adj : sourd-muet, ivre-mort

# Wooclap


1 Allez sur wooclap.com
2 Entrez le code d'événement dans le bandeau supérieur
Code d'événement
EZLXLK
Activer les réponses par SMS

# On se réveille (wooclap)

**Exercice**

Quel est le pluriel de *porte-manteau* ? de *savoir-vivre* ? de *sourd-muet* ?

# Certaines langues composent plus que d'autres...


https://www.youtube.com/watch?v=gG62zay3kck

# Agglutination

*Réunion en une seule unité de deux ou plusieurs termes originairement distincts (Wikipédia)*

- aujourd'hui : au jour d'hui

# Agglutination

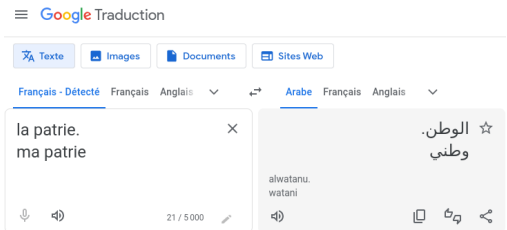
*Réunion en une seule unité de deux ou plusieurs termes originairement distincts (Wikipédia)*

- aujourd'hui : au jour d'hui
- alentour : à l'entour

# Agglutination

*Réunion en une seule unité de deux ou plusieurs termes originairement distincts (Wikipédia)*

- aujourd'hui : au jour d'hui
- alentour : à l'entour
- en arabe le déterminant est "collé" au nom :



# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)
- *epäjärjestelmällistyttämättömyydelläänsäkäänköhän* :

# Certaines langues agglutinent plus que d'autres...

Exemples en finnois :

- *Järjestelmällinen* : organisé
- *järjestelmällistetty* : conçu pour être organisé
- *epäjärjestelmällistetty* : désorganisé ou pas conçu pour être organisé
- *Epäjärjestelmällistämättömyy* : négation et substantivation
- *Epäjärjestelmällistyttämättömyydellänsä* : l'objet dénoté par le substantif appartient à quelqu'un (génitif)
- *Epäjärjestelmällistyttämättömyydellänsäkään* : nouvelle négation
- *epäjärjestelmällistyttämättömyydellänsäkäänkö* : ajout d'un suffixe dénotant l'interrogation
- *epäjärjestelmällistyttämättömyydellänsäkäänköhän* : ajout d'une emphase (c'est moi qui ...)
- *epäjärjestelmällistyttämättömyydelläänsäkäänköhän* :

  *The reverse of the reverse of something abstract that is made to be unorganised, which is owned by someone, and is one of the two or more (possibly similar) attributes that have a negative atmosphere or lack of something, and we doubt if it is it at the same time that we ensure that it truly is.*

# La polysynthèse : une agglutination extrême

Combinaison de :

- nombreux morphèmes
- de manière pas nécessairement linéaire (pas bout à bout)
- pour former un mot-phrase

Notamment dans certaines langues inuits ou aborigènes d'Australie

# Exemple de langue polysynthétique : le yupik

Mangteghaghllangllaghyugtukut
house-big-to.make-to.want.to-IND.INTR-1PL
'We want to make a big house.'

Bruno Guillaume, présenté à LIFT2, avec son accord

Phrases et tours de parole

Rien que des mots, toujours des mots

On fléchit, on dérive, on compose, on agglutine : la morphologie

Pour finir

CQFR : Ce Qu'il Faut Retenir


- mot vs token vs unité lexicale
- flexion vs dérivation vs composition vs agglutination

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., and Tsvetkov, Y. (2023).
Do all languages cost the same ? tokenization in the era of commercial language models.
In Bouamor, H., Pino, J., and Bali, K., editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9904–9923, Singapore. Association for Computational Linguistics.

de Saussure, F. (1916).
Cours de linguistique générale.
Payot, Paris.

Sagot, B. (2018).
Informatiser le lexique.
Habilitation à diriger des recherches, Sorbonne Université.