Speech and language processing

Chloé Clavel, Rachel Bawden

Benoît Sagot, Emmanuel Dupoux, Maxime Poli

MVA – Speech and Language processing – Class #1 – 15th January 2026

Course logistics

All information, including (possibly last-minute) updates are available on the class's GitHub site:

https://github.com/rbawden/MVA_2026_SNLP

- Classes will be given by **Chloé Clavel (Inria, course manager)**, **Rachel Bawden (Inria, course manager)**, **Benoît Sagot** (Inria), **Emmanuel Dupoux** (EHESS, ENS), and **Maxime Poli** (ENS)

- Contact: mva.speech.language@gmail.com

Evaluation

- **Quizzes (40% of the total grade)**
  - Online questionnaire (google form) after each class (except this one and the last one), 30 minutes to complete it
    - forms submitted after the deadline will be automatically rejected and graded as zero
  - Incorrect answer = half the points (allows us to check attendance to the class)
  - The best 6 grades out of the 7 quizzes will be used for the average (=> you can skip/miss one quiz out of 7)

Evaluation

- **Quizzes (40% of the total grade)**
  - Online questionnaire (google form) after each class (except this one and the last one), 30 minutes to complete it
    - forms submitted after the deadline will be automatically rejected and graded as zero
  - Incorrect answer = half the points (allows us to check attendance to the class)
  - The best 6 grades out of the 7 quizzes will be used for the average (=> you can skip/miss one quiz out of 7)
- **Projects (60% of the total grade) – cf. next slide**

Evaluation

- **Quizzes (40% of the total grade)**
  - Online questionnaire (google form) after each class (except this one and the last one), 30 minutes to complete it
    - forms submitted after the deadline will be automatically rejected and graded as zero
  - Incorrect answer = half the points (allows us to check attendance to the class)
  - The best 6 grades out of the 7 quizzes will be used for the average (=> you can skip/miss one quiz out of 7)
- **Projects (60% of the total grade) – cf. next slide**
- **There will be no second session (“pas de rattrapage”)**
  - Total grade < 10/20 = failure to validate the course
  - For exceptional circumstances, please come and see us

Evaluation

Evaluation

- **Projects (60% of the total grade)**
  - Based on a recent paper+code, to be chosen within a list that will be shared by the 23rd January
    - You will express individual preferences, we will define the groups based on them (~3 students/group, random assignment)
    - Individual choice of project topic due on Fri 30th January
    - We will share with you the composition of the groups by the 2nd February
  - Objective: replicate the paper and extend it with (an) additional experiment(s)
  - **One-page outline due on Fri 20th February end of the day**
    - You will have a quick meeting with us or one of our PhD student to get feedback on your one-pager
  - **Final report (maximum 4 pages) + GitHub repo with your code due on Fri 26th March end of the day**
    - Strictly enforced deadline (1/24th point /20 substracted every late hour). Points subtracted if over 4 pages.
  - **Oral presentation** (10min presentation + 5min questions) between 30th March and 9th April

Course overview 1/2 (check the GitHub site)

All classes will take place in the Salle des Actes, ENS, 45 rue d'Ulm

1. Thu 15th January (2pm): **Introduction + Language modelling 1** (B. Sagot)
2. Thu 22nd January (2pm): **Language modelling 2** (B. Sagot)
3. Thu 29th January (2pm): **Sentiment analysis** (C. Clavel)
    - Fri 30th January: individual choice of project topic due, groups' composition announced within the next few days
4. Thu 5th February (2pm): **Machine translation** (R. Bawden)
5. Thu 12th February (2pm): **Speech processing 1: acoustic models for ASR** (M. Poli)
    - Fri 20th February: project one-pager due

Slots are 3 hours long. Most classes will last approx. 2 to 2.5 hours, leaving time for a quiz and time for questions

Course overview 2/2 (check the GitHub site)

All classes will take place in the Salle des Actes, ENS, 45 rue d'Ulm

6. Thu 26th February (2pm): **Speech processing 2** (E. Dupoux)
7. Thu 5th March (2pm): **Speech processing 3** (E. Dupoux)
8. Thu 12th March (2pm): **Conversational systems** (C. Clavel)
9. Thu 19th March (2pm): ***Guest lecture*** (G. Synnaeve, META)
    - Thu 26th March: final report due
    - Between 30th March and 9th April: project defences (precise dates to be confirmed)

Slots are 3 hours long. Most classes will last approx. 2 to 2.5 hours, leaving time for a quiz and time for questions

Introduction to speech and language processing

Benoît Sagot

MVA – Speech and Language processing – Class #1.1 – 16 January 2025

AI, NLP and speech processing

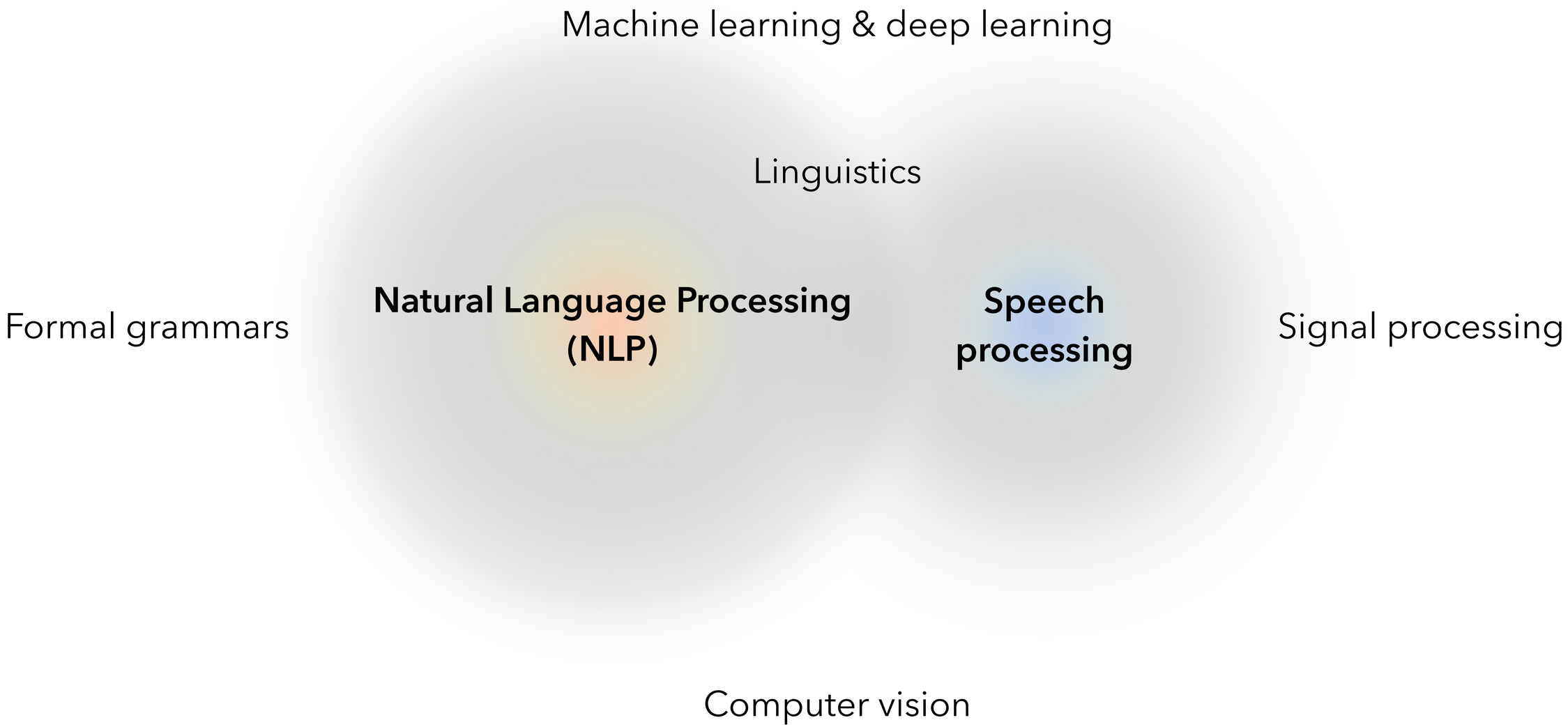
• On the left side of the image, a young woman is speaking in a microphone. On the right side, a young man is reading a message on his laptop, which is the translation of the woman's message. Between both sides of the image, represent the fact that the two people are very far away from one another and connected through the Internet.
• Make the image square. • Add a touch of steampunk style in the image.
DALL-E 3 (via ChatGPT, GPT-4, 04/02/2024)

Artificial intelligence (AI)

- **AI** = research and engineering field whose goal is to **computationally reproduce (or imitate) behaviours that normally require human intelligence**
  - Using language is one of them

Source : https://www.pexels.com/fr-fr/photo/femme-jouer-table-technologie-8438940/ (Pavel Danilyuk, free to use)

NLP, speech processing and adjacent fields



Edited text

User-generated content

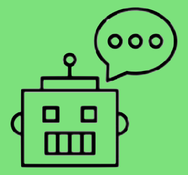
Speech transcriptions

Natural Language Processing (NLP)

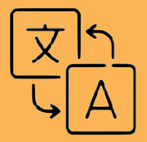
Core tasks :

- Morphosyntactic tagging
- Syntactic parsing
- Semantic parsing
- etc.

Analysis



Generation



Transformation

Interaction

Analysis

Generation

Transformation

Interaction

Spelling correction

Artificial intelligence (AI) is the inteligence of machines or software, as opposed to the intelligence of living beings, primarily of humans. It is a fields of study in computer science that develops and studies intelligen nines may be called AIs.

field of study

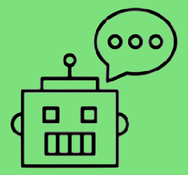
Ignorer

Grammaire...

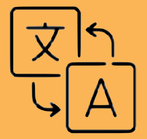
Recherche intelligente... ^⌘L

Microsoft Word

Analysis



Generation



Transformation

Interaction

Spelling correction

Hate speech detection

Analysis

Generation

Transformation

Interaction

Spelling correction

Hate speech detection

Information extraction

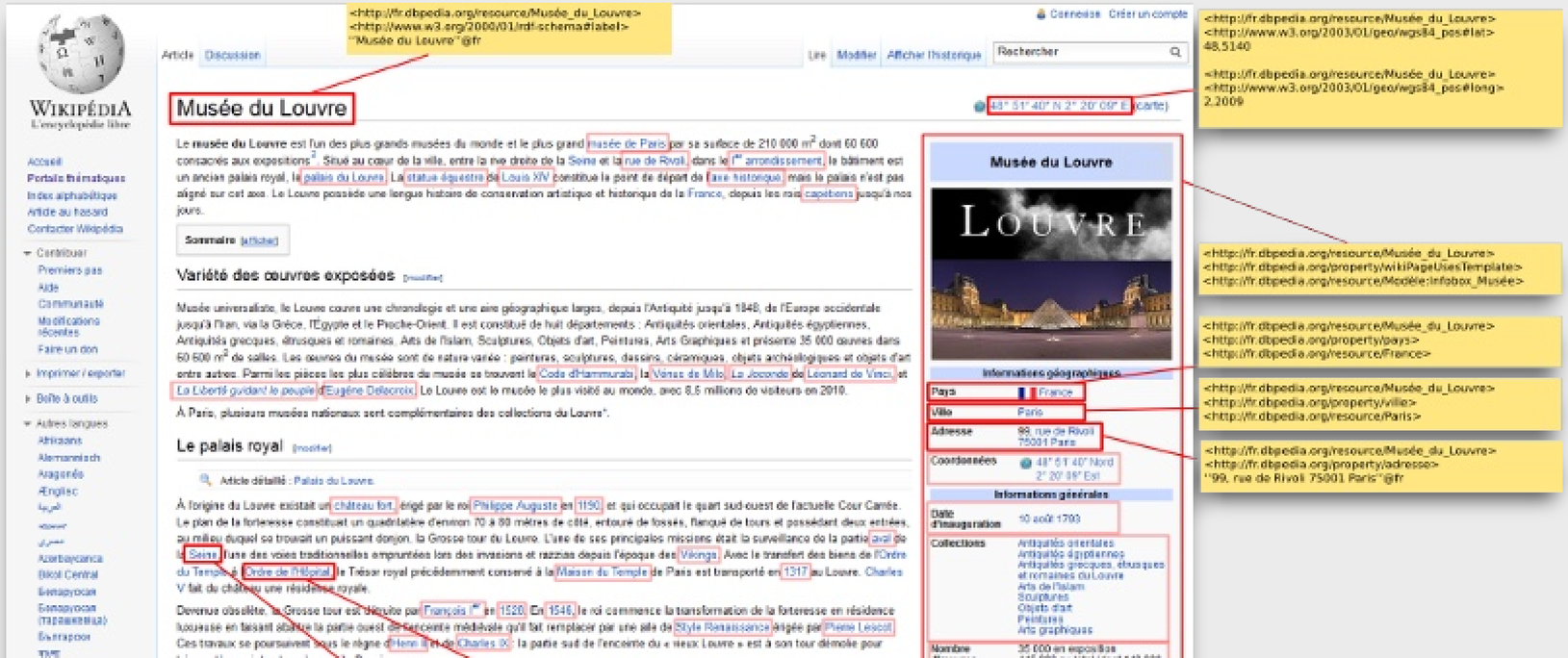


Figure: DBpedia.fr - Inria (WIMMICS)

Analysis

Spelling correction

Hate speech detection

Information extraction

Generation

Data-to-text

Transformation

Interaction

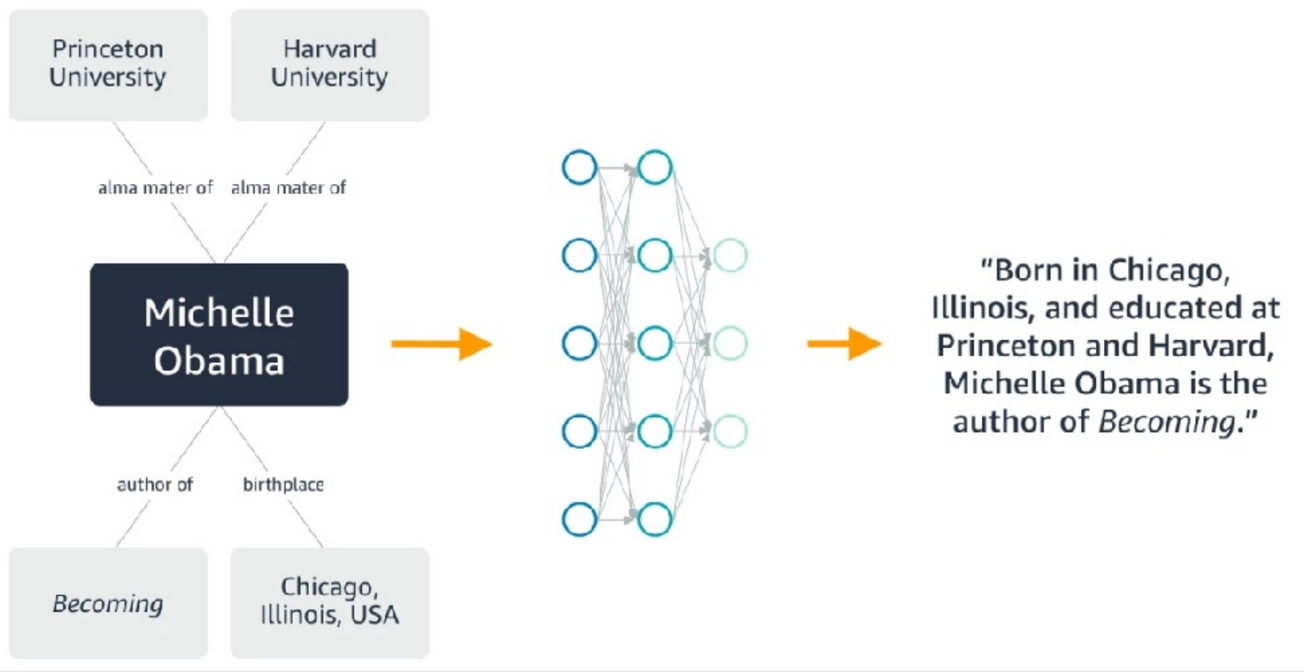


Figure: Amazon Science (Glynis Condon)

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text		
Hate speech detection	Image captioning		
Information extraction			

Tabby cat wide-open with green eyes

Photo BS

Analysis

Spelling correction

Hate speech detection

Information extraction

Generation

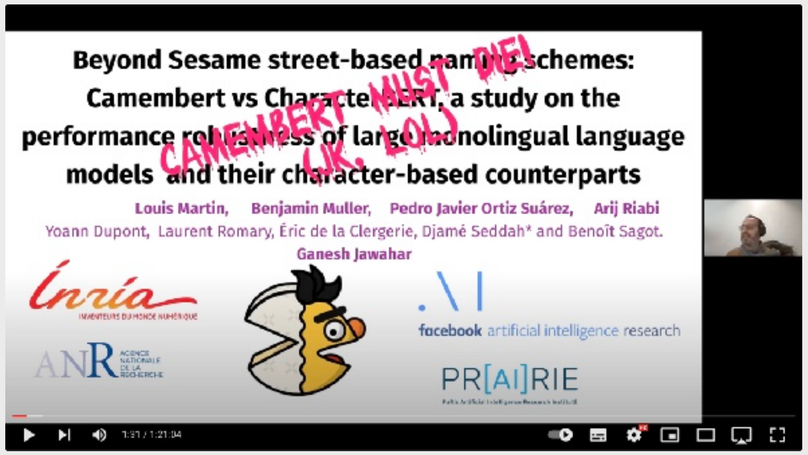
Data-to-text

Image captioning

Video summarisation

Transformation

Interaction



EnfinBref.io

Copiez simplement le lien d'une vidéo YouTube pour en obtenir le résumé, généré par l'intelligence artificielle.
Le résumé est généré en français 🇫🇷, même pour les vidéos dans d'autres langues.

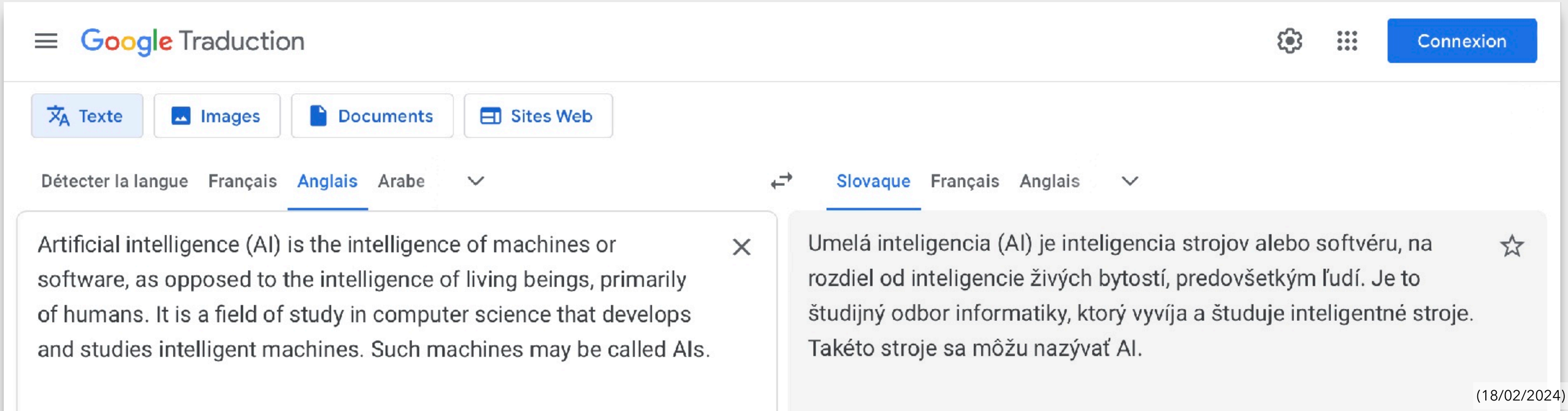
Résumé de la vidéo CamemBERT must die! (jk,lol) 🇫🇷

Tous les résumés sont générés automatiquement par l'IA, des erreurs peuvent s'y trouver.

Dans cette vidéo, les chercheurs présentent les modèles de langue et discutent de différentes techniques pour améliorer leurs performances. Ils présentent le modèle d'apprentissage "Camembert" et abordent des sujets tels que l'ajout de tâches de connaissances et l'entraînement incrémental des couches du modèle. Ils soulignent également les défis liés à la diversité linguistique et proposent une approche basée sur les caractères pour améliorer la robustesse des modèles. La prochaine séance est annoncée dans quatre semaines.

Retrieved in November 2023. The site is now (25/02/2024) password-protected.

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	
Hate speech detection	Image captioning		
Information extraction	Video summarisation		



Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	
Hate speech detection	Image captioning	Text summarisation	
Information extraction	Video summarisation		

Artificial intelligence (AI) is the intelligence of machines or software, as opposed to the intelligence of living beings, primarily of humans. It is a field of study in computer science that develops and studies intelligent machines. Such machines may be called AIs.

Artificial intelligence (AI) refers to machine or software intelligence, distinct from human or living beings' intelligence, studied within computer science.

Summary produced using GPT-4 (29/02/2024)

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	
Hate speech detection	Image captioning	Text summarisation	
Information extraction	Video summarisation	Text simplification	

ATTESTATION DE DÉPLACEMENT DÉROGATOIRE

En application de l'article 3 du décret du 23 mars 2020 prescrivant les mesures générales nécessaires pour faire face à l'épidémie de Covid19 dans le cadre de l'état d'urgence sanitaire

Je soussigné(e),

Mme/M. :

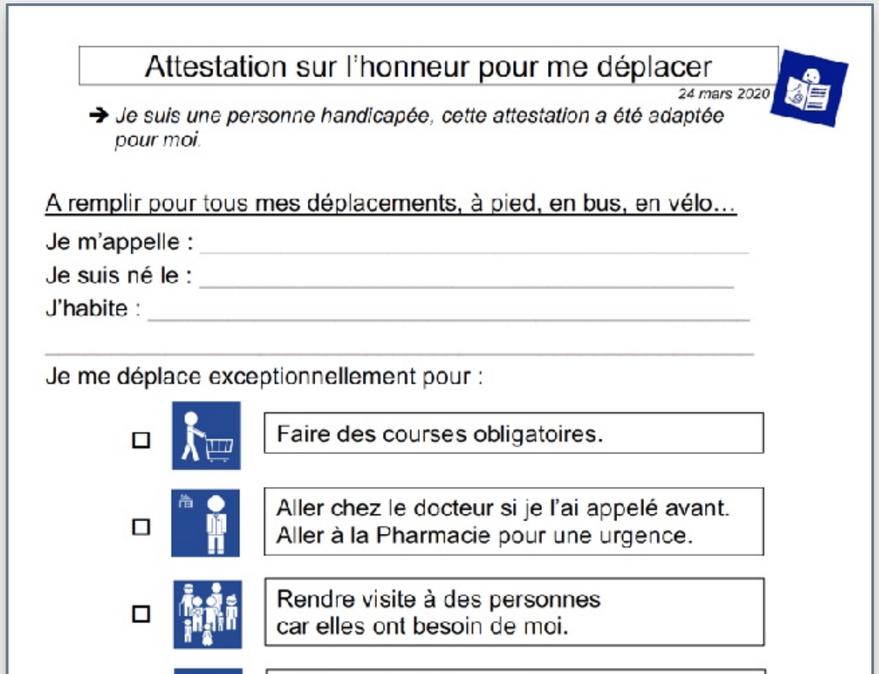
Né(e) le :

À :

Demeurant :

certifie que mon déplacement est lié au motif suivant (cocher la case) autorisé par l'article 3 du décret du 23 mars 2020 prescrivant les mesures générales nécessaires pour faire face à l'épidémie de Covid19 dans le cadre de l'état d'urgence sanitaire[1] :

- ☐ Déplacements entre le domicile et le lieu d'exercice de l'activité professionnelle, lorsqu'ils sont indispensables à l'exercice d'activités ne pouvant être organisées sous forme de télétravail ou déplacements professionnels ne pouvant être différés[2].
- ☐ Déplacements pour effectuer des achats de fournitures nécessaires à l'activité professionnelle et des achats de première nécessité[3] dans des établissements dont les activités demeurent autorisées (liste sur gouvernement.fr).
- ☐ Consultations et soins ne pouvant être assurés à distance et ne pouvant être différés ; consultations et soins des patients atteints d'une affection de longue durée.
- ☐ Déplacements pour motif familial impérieux, pour l'assistance aux personnes vulnérables ou la garde d'enfants.



Attestation sur l'honneur pour me déplacer

24 mars 2020

➔ *Je suis une personne handicapée, cette attestation a été adaptée pour moi.*

A remplir pour tous mes déplacements, à pied, en bus, en vélo…

Je m'appelle : _____

Je suis né le : _____

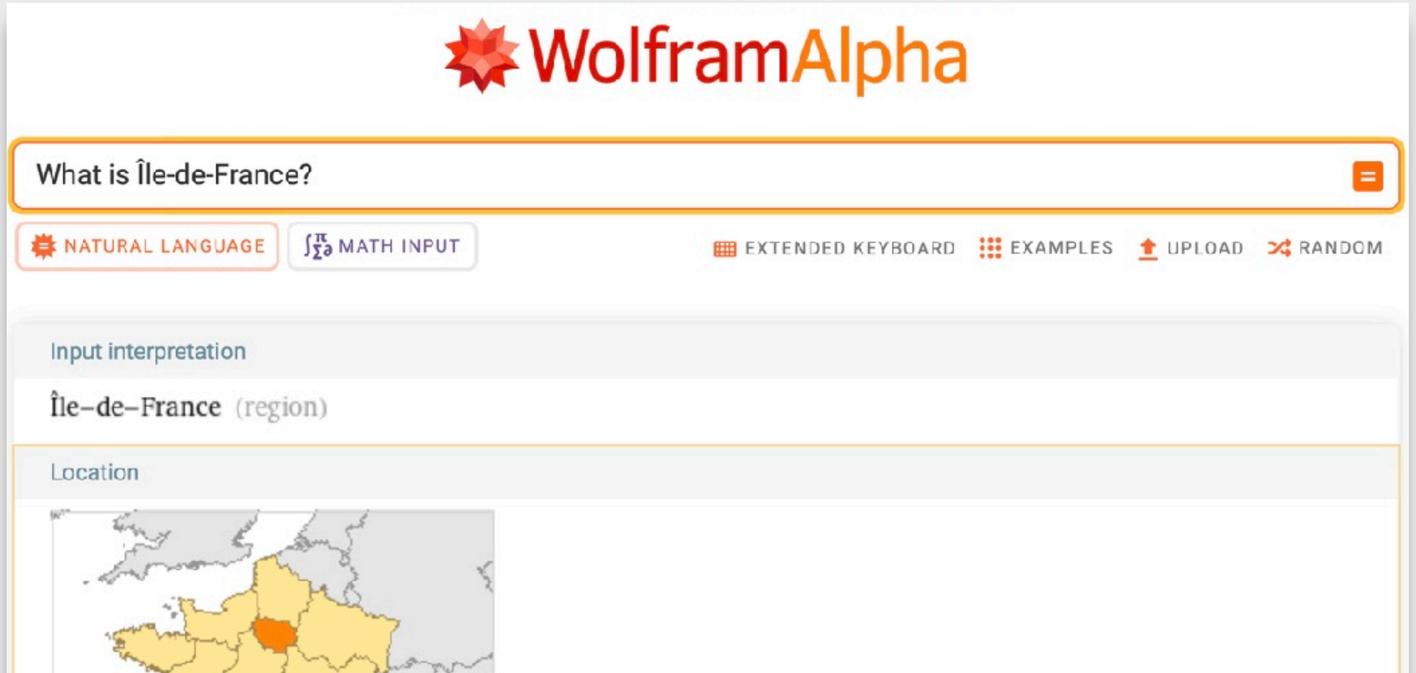
J'habite : _____

_____

Je me déplace exceptionnellement pour :

- ☐ Faire des courses obligatoires.
- ☐ Aller chez le docteur si je l'ai appelé avant. Aller à la Pharmacie pour une urgence.
- ☐ Rendre visite à des personnes car elles ont besoin de moi.

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	
Information extraction	Video summarisation	Text simplification	



WolframAlpha (18/02/2024)

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	

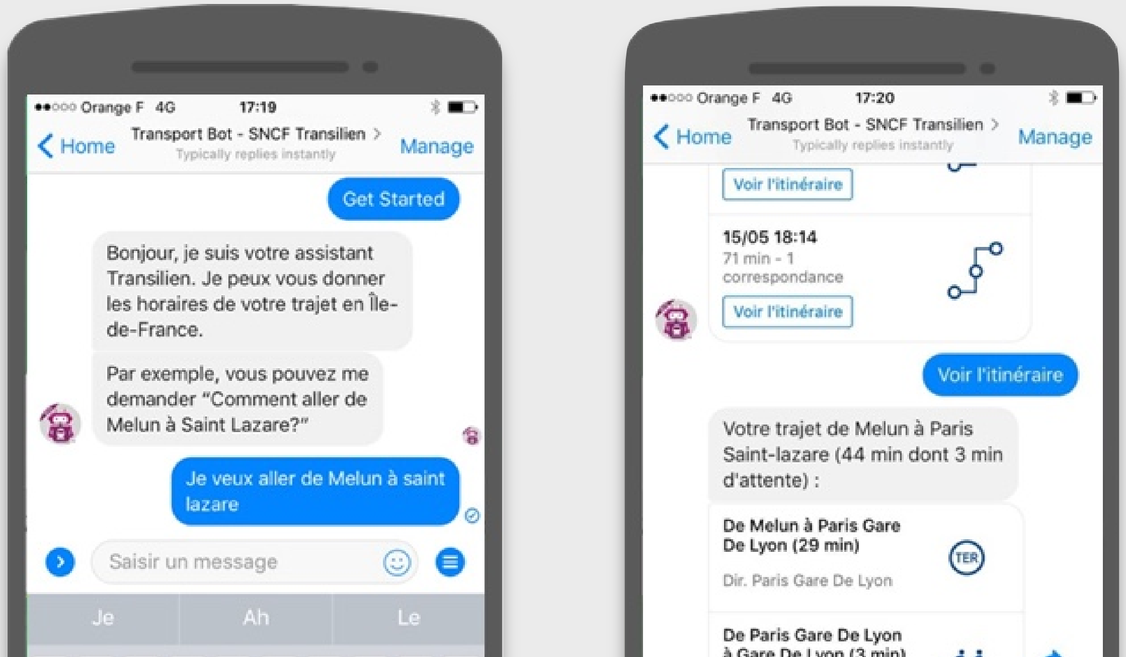
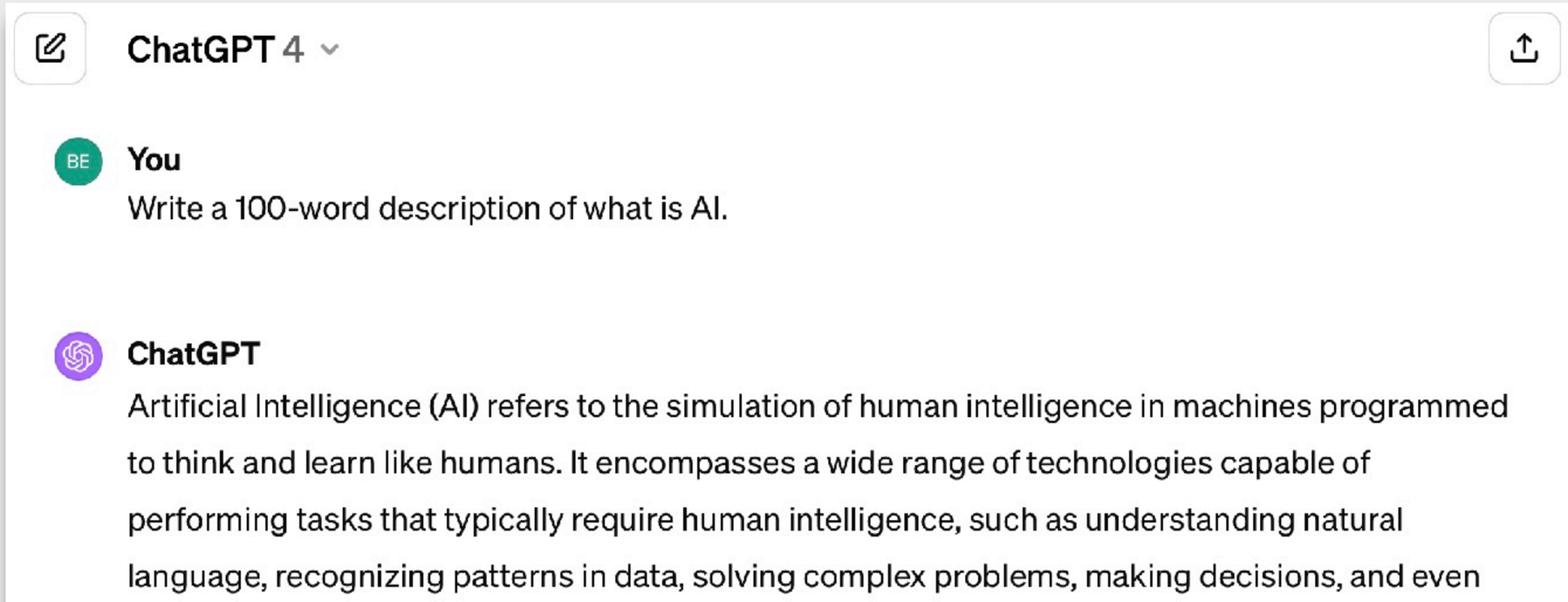
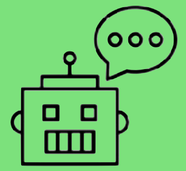
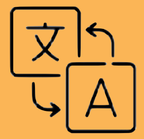


Illustration: SNCF

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots



ChatGPT (GPT-4, 29/02/2024)

 **Analysis**	 **Generation**	 **Transformation**	 **Interaction**
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots

Application to other fields

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots

Application to other fields

Linguistics

- Modelling languages
- Modelling their evolution
- Corpus linguistics

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots

Application to other fields

Linguistics

- Modelling languages
- Modelling their evolution
- Corpus linguistics

Digital humanities

- Analysis of large text corpora
- E.g. archives, computational philology

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots

Application to other fields

Linguistics

- Modelling languages
- Modelling their evolution
- Corpus linguistics

Digital humanities

- Analysis of large text corpora
- E.g. archives, computational philology

Law

- Legal document analysis
- Case law research
- Plagiarism detection

Analysis	Generation	Transformation	Interaction
Spelling correction	Data-to-text	Machine translation	Question answering
Hate speech detection	Image captioning	Text summarisation	Specialised chatbots
Information extraction	Video summarisation	Text simplification	General chatbots

Application to other fields

Linguistics

- Modelling languages
- Modelling their evolution
- Corpus linguistics

Digital humanities

- Analysis of large text corpora
- E.g. archives, computational philology

Law

- Legal document analysis
- Case law research
- Plagiarism detection

Health

- Medical record analysis
- Scientific publication analysis
- Early diagnosis

High-quality voice data

Voice data in noisy contexts

Voice over the phone

Speech Processing

Analysis

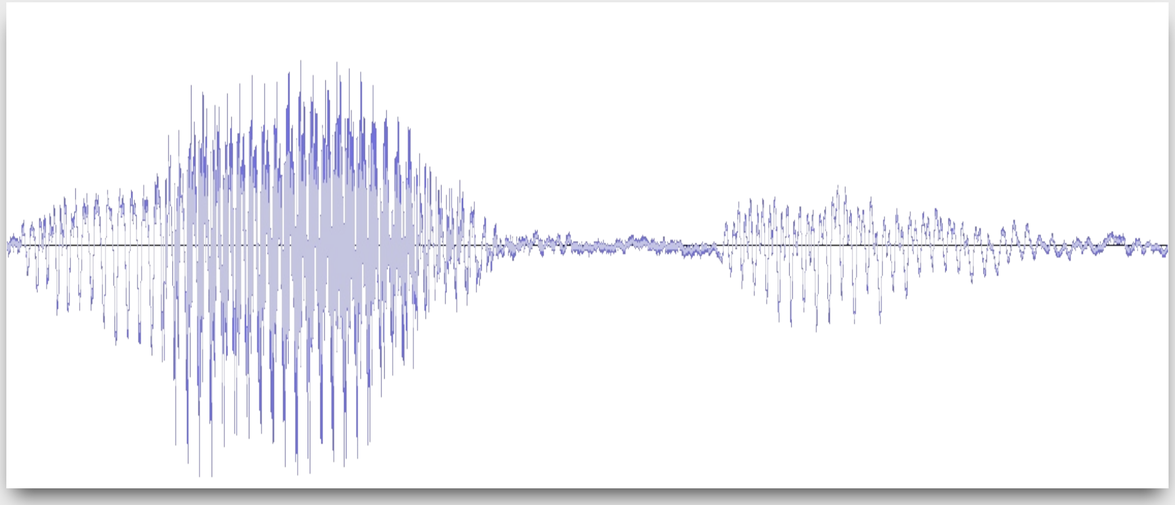
Generation

Transformation

Interaction

Analysis | **Generation** | **Transformation** | **Interaction**

Automatic speech recognition



Martin

Source: Wikipedia (MartinThoma)

Analysis

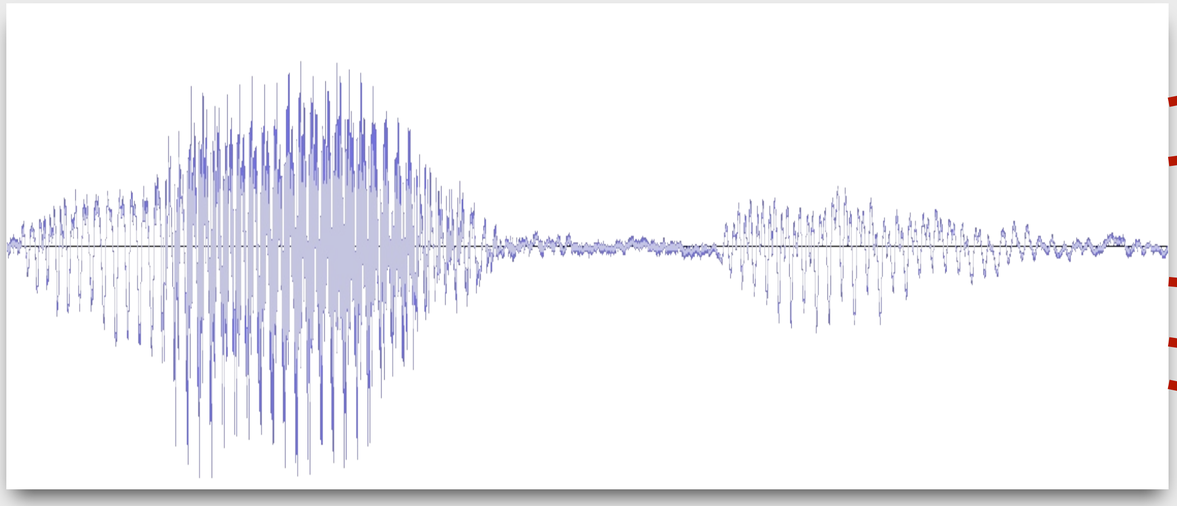
Generation

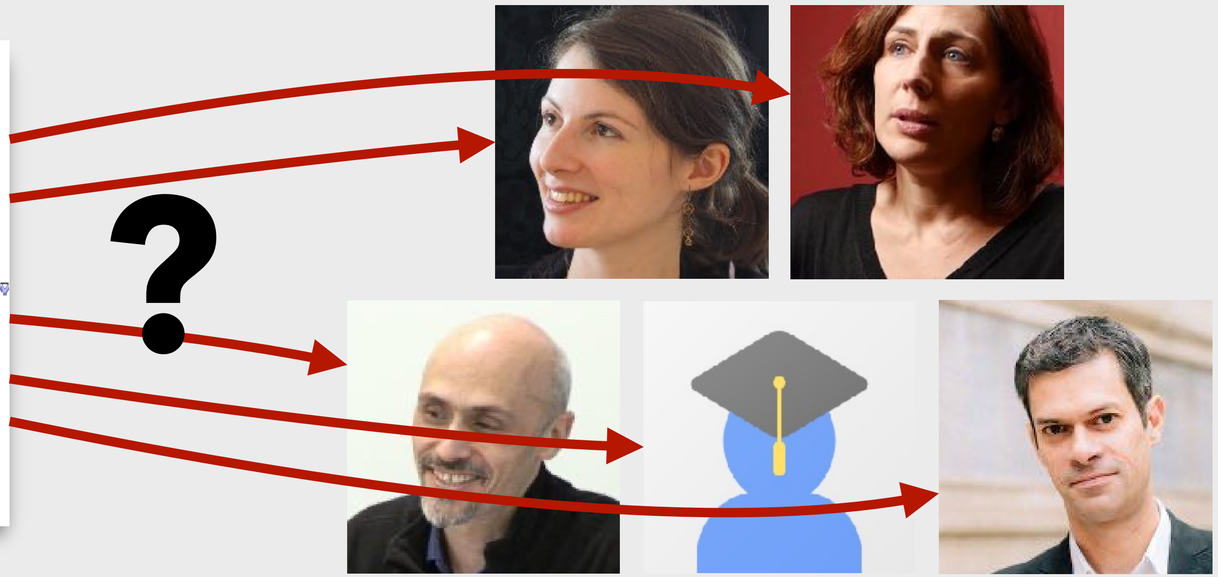
Transformation

Interaction

Automatic speech recognition

Speaker identification





Analysis

Generation

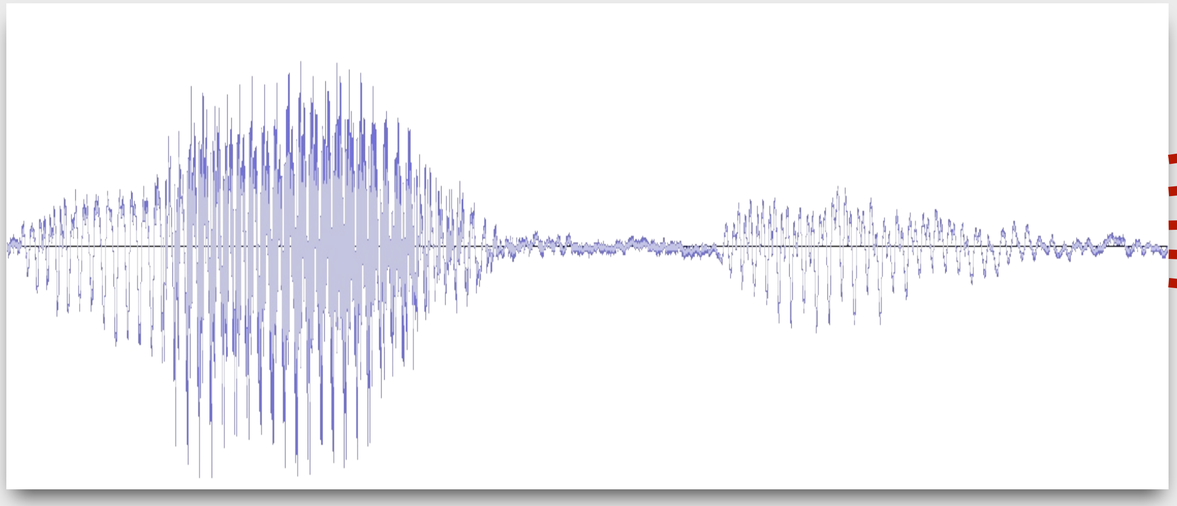
Transformation

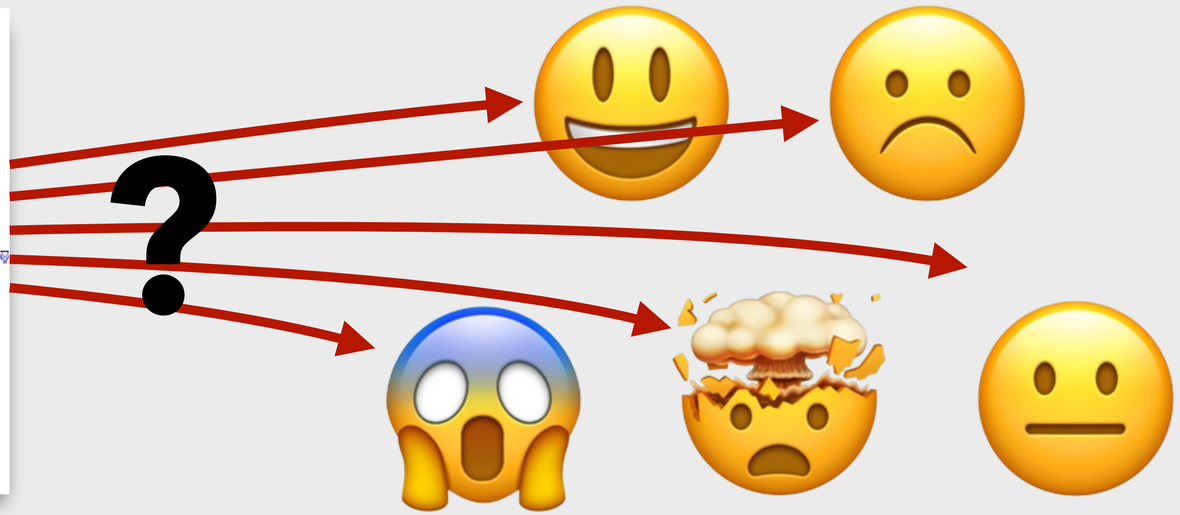
Interaction

Automatic speech recognition

Speaker identification

Emotion identification





Analysis

Automatic speech recognition

Speaker identification

Emotion identification

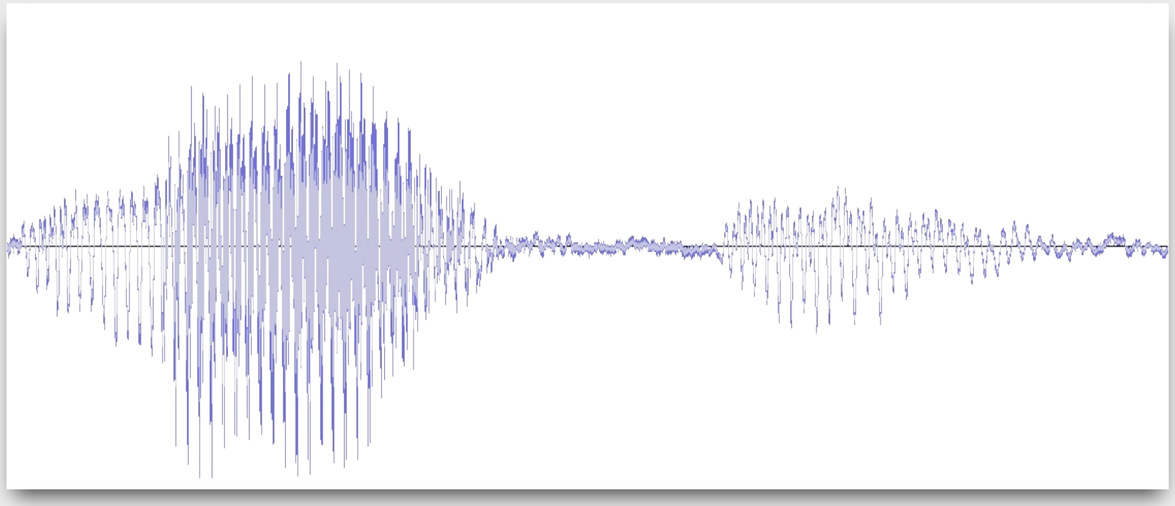
Generation

Speech synthesis

Transformation

Interaction

Martin



Analysis

- Automatic speech recognition
- Speaker identification
- Emotion identification

Generation

- Speech synthesis
- Screen readers

Transformation

Interaction

Source: Wikipedia (Apple)

Analysis	Generation	Transformation	Interaction
Automatic speech recognition	Speech synthesis	Speech-to-speech translation	
Speaker identification	Screen readers		
Emotion identification			

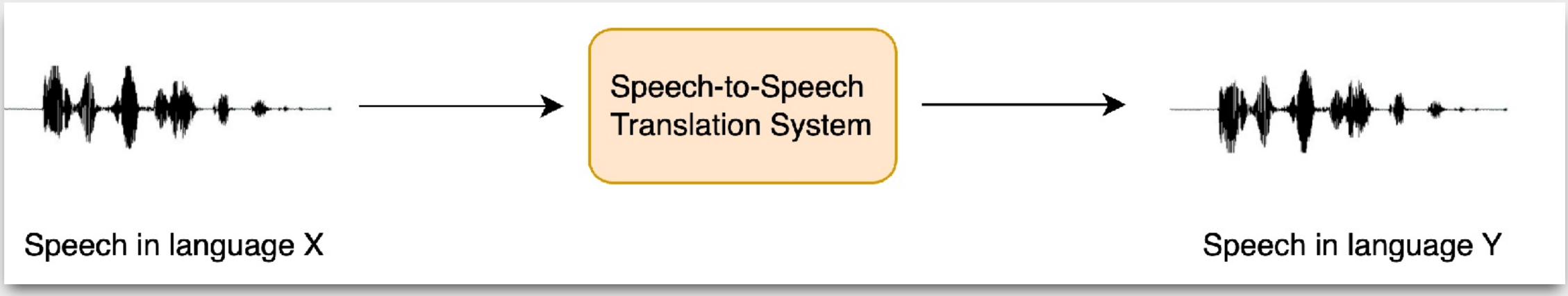


Figure from a web course published by Hugging Face

Analysis

- Automatic speech recognition
- Speaker identification
- Emotion identification

Generation

- Speech synthesis
- Screen readers

Transformation

- Speech-to-speech translation
- Emotion conversion

Interaction

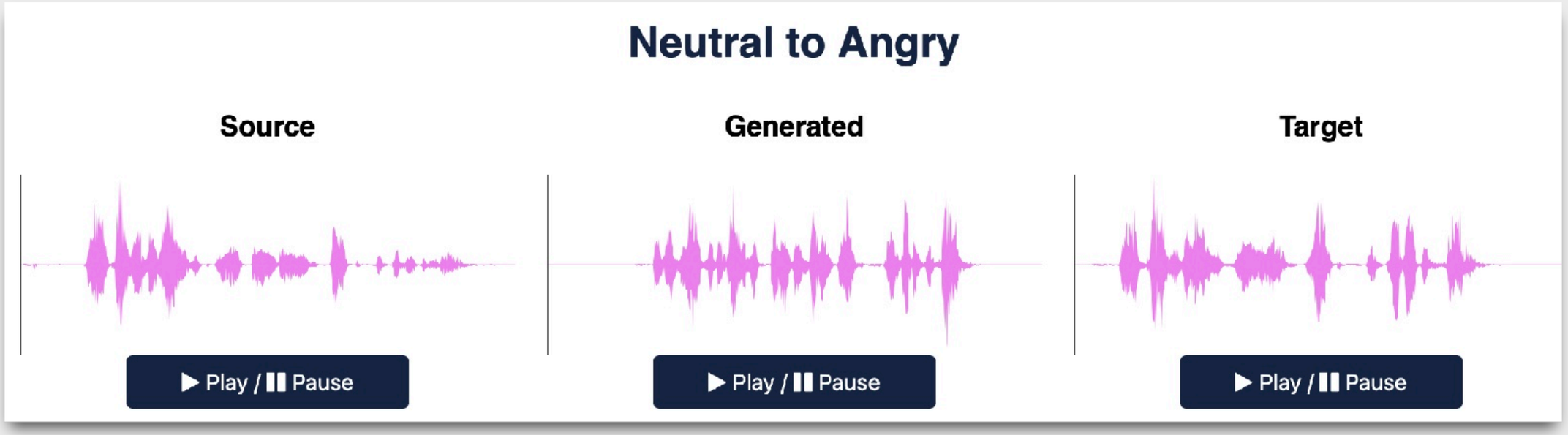


Figure from Kreuk *et al.* (2022)'s GitHub page

Analysis	Generation	Transformation	Interaction
Automatic speech recognition	Speech synthesis	Speech-to-speech translation	
Speaker identification	Screen readers	Emotion conversion	
Emotion identification		Speaker modification	

Analysis	Generation	Transformation	Interaction
Automatic speech recognition	Speech synthesis	Speech-to-speech translation	Voice chatbots
Speaker identification	Screen readers	Emotion conversion	
Emotion identification		Speaker modification	

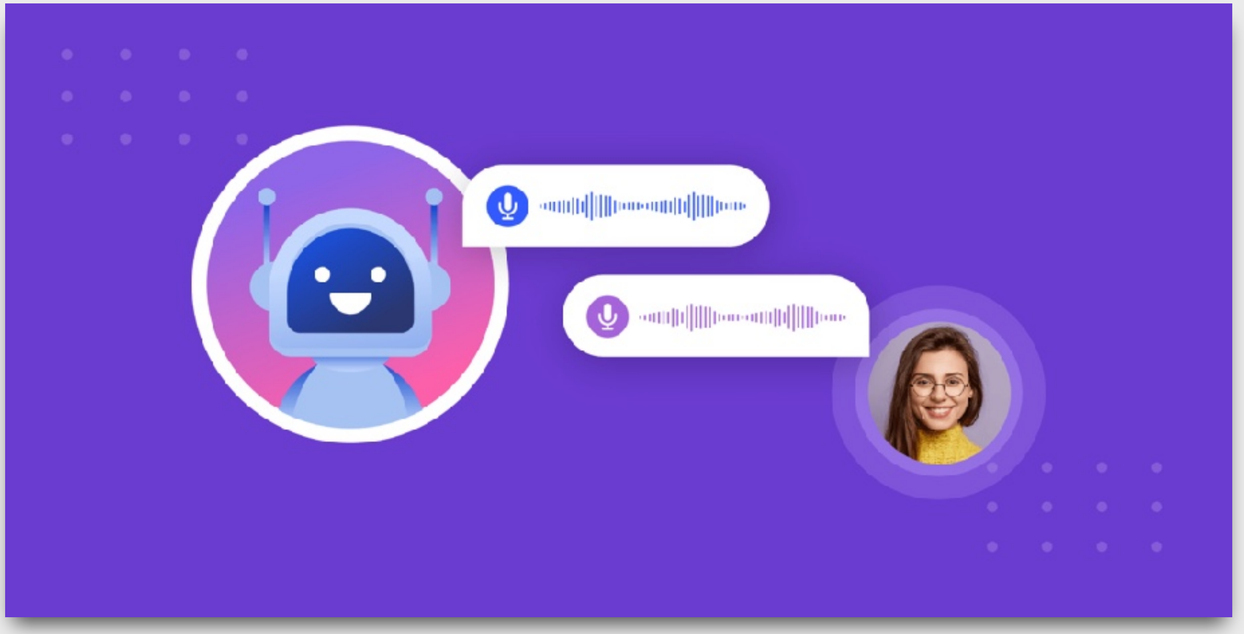


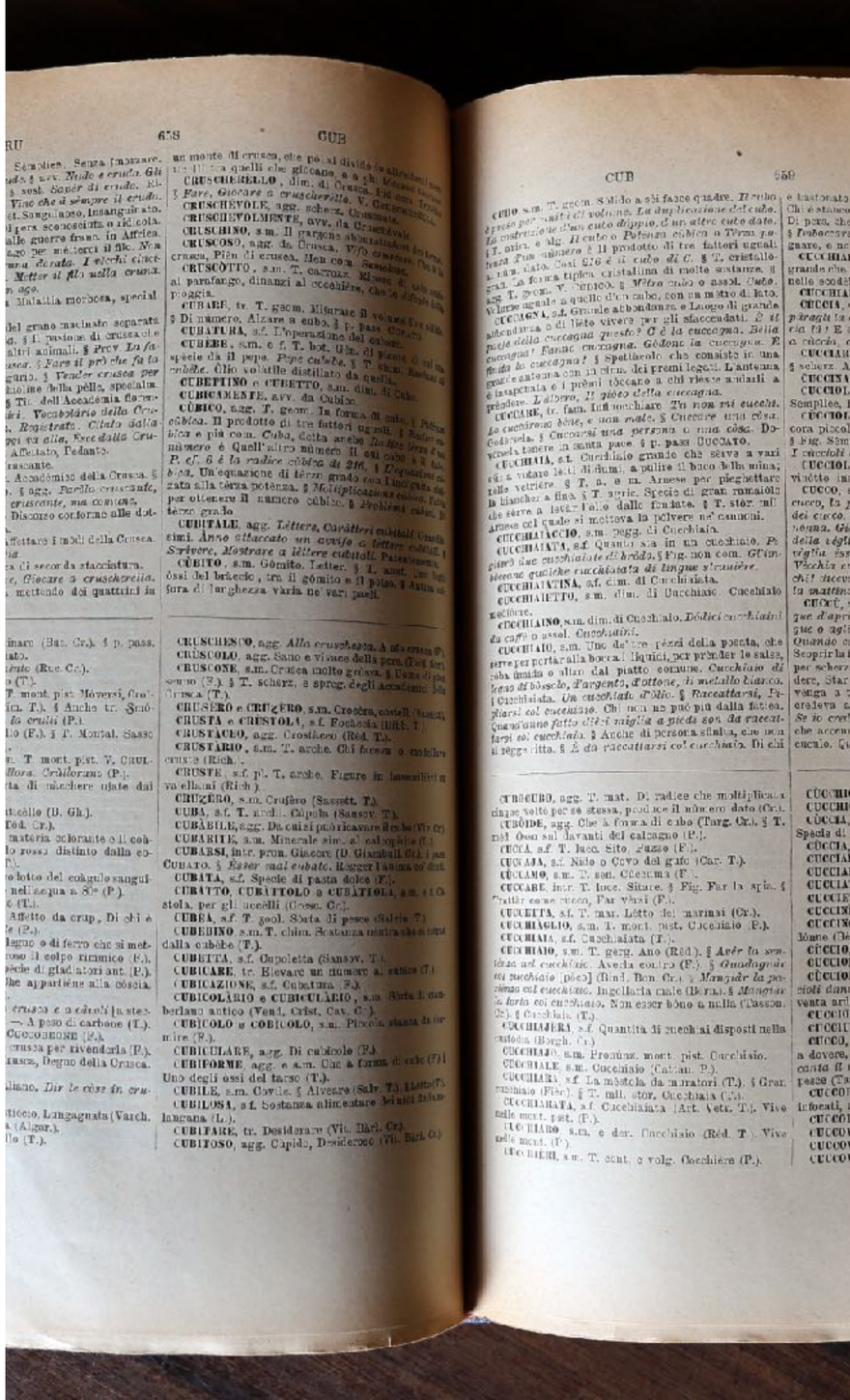
Figure from verloop.io

Analysis	Generation	Transformation	Interaction
Automatic speech recognition	Speech synthesis	Speech-to-speech translation	Voice chatbots
Speaker identification	Screen readers	Emotion conversion	Virtual assistants
Emotion identification		Speaker modification	

Sources: Apple, Amazon

NLP, speech processing and linguistics

- NLP and speech processing are dedicated to **processing text/speech data**
  - **Linguistics is the science exploring such data**
- **The relationship between speech processing/NLP and linguistics has evolved over time**
  - Symbolic approaches: formalisation, description (grammars, lexicons)
  - Machine-learning-based approaches: linguistic annotation in corpora
- Any field relying on machine learning **can take advantage of a better understanding of the data at hand**
  - Joint terminology
  - Better anticipation and understanding of challenges
  - Linguistic data is structured, complex, and discrete
    - Understanding such data might be even more useful in NLP and speech processing than in other fields

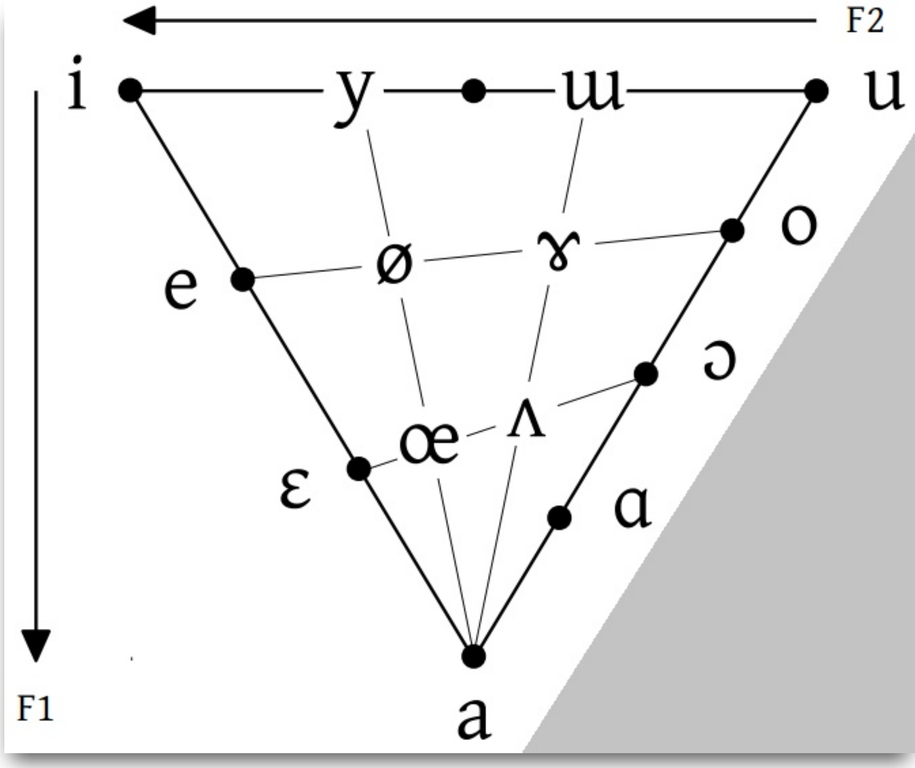


A brief introduction to linguistics

• *Draw linguistics, steampunk style.*
DALL-E 3 (via ChatGPT, GPT-4, 07/12/2023)

Phonology

How a language's sounds are organised



- **Phonetics** = study of sounds in language in general
  - Atomic unit = phone, displayed between square brackets: [r]
- **Phonology** = in a given language, how phones are clustered into sets of interchangeable (although often contextually dependent) sets
  - Atomic unit = phoneme, displayed between slashes: /r/
- **Phone–phoneme relationships are language-dependent**
  - In French, /r/ can be realised as [ʁ],[ʁ̞],[χ],[ʀ],[r],[ɾ] (the last two are perceived as dialectal)
  - In Abkhaz, [ʁ], [χ], and [r] are 3 different phonemes that can be written /ʁ/, /χ/, and /r/
  - In French, the phoneme closest to English's phonemes /ɪ/ and /iː/ is the same one: /i/

Figure: Wikipedia

Morphology

How words are built

- Inflectional morphology
  - {*mange*, *mangiez*, *mangerions*, *mangeaient*, …} : « manger » ‘eat’
  - Lemma, inflexion class, morphological/ morphosyntactic features
- Derivational morphology
  - *brav-itude*, *bio*+*terror-isme*/*-iste*, *skype*+*(e)r*
- Morphemes, morphs

Conjugaison de **manger**, *verbe du 1er groupe*, *conjugué avec l'auxiliaire avoir*.

Formes du participe passé

	Singulier	Pluriel
Masculin	mangé \mɑ̃.ʒe\	mangés \mɑ̃.ʒe\
Féminin	mangée \mɑ̃.ʒe\	mangées \mɑ̃.ʒe\

Modes impersonnels

Conjugaison en français
manger
Verbe du premier groupe, conjugué comme {{fr-conj-1-ger}}

Mode	Présent		Passé	
Infinitif	manger	\mɑ̃.ʒe\	avoir mangé	\a.vwaʁ mɑ̃.ʒe\
Gérondif	en mangeant	\ɑ̃ mɑ̃.ʒɑ̃\	en ayant mangé	\ɑ̃.n_ɛ.jɑ̃ mɑ̃.ʒe\
Participe	mangeant	\mɑ̃.ʒɑ̃\	mangé	\mɑ̃.ʒe\

Indicatif

Présent		Passé composé	
je mange	\ʒə mɑ̃ʒ\	j'ai mangé	\ʒ_e mɑ̃.ʒe\
tu manges	\ty mɑ̃ʒ\	tu as mangé	\ty a mɑ̃.ʒe\
il/elle/on mange	\[il/ɛl/ɔ̃] mɑ̃ʒ\	il/elle/on a mangé	\[i.l/ɛ.l/ɔ̃.n]_a mɑ̃.ʒe\
nous mangeons	\nu mɑ̃.ʒɔ̃\	nous avons mangé	\nu.z_a.vɔ̃ mɑ̃.ʒe\
vous mangez	\vu mɑ̃.ʒe\	vous avez mangé	\vu.z_a.ve mɑ̃.ʒe\
ils/elles mangent	\[il/ɛl] mɑ̃ʒ\	ils/elles ont mangé	\[i/ɛ]l.z_ɔ̃ mɑ̃.ʒe\
Imparfait		**Plus-que-parfait**	
je mangeais	\ʒə mɑ̃.ʒɛ\	j'avais mangé	\ʒ_a.vɛ mɑ̃.ʒe\
tu mangeais	\ty mɑ̃.ʒɛ\	tu avais mangé	\ty a.vɛ mɑ̃.ʒe\
il/elle/on mangeait	\[il/ɛl/ɔ̃] mɑ̃.ʒɛ\	il/elle/on avait mangé	\[i.l/ɛ.l/ɔ̃.n]_a.vɛ mɑ̃.ʒe\
nous mangions	\nu mɑ̃.ʒjɔ̃\	nous avions mangé	\nu.z_a.vjɔ̃ mɑ̃.ʒe\
vous mangiez	\vu mɑ̃.ʒje\	vous aviez mangé	\vu.z_a.vje mɑ̃.ʒe\
ils/elles mangeaient	\[il/ɛl] mɑ̃.ʒɛ\	ils/elles avaient mangé	\[i/ɛ]l.z_a.vɛ mɑ̃.ʒe\
Passé simple		**Passé antérieur**	
je mangeai	\ʒə mɑ̃.ʒe\	j'eus mangé	\ʒ_y mɑ̃.ʒe\
tu mangeas	\ty mɑ̃.ʒa\	tu eus mangé	\ty y mɑ̃.ʒe\
il/elle/on mangea	\[il/ɛl/ɔ̃] mɑ̃.ʒa\	il/elle/on eut mangé	\[i.l/ɛ.l/ɔ̃.n]_y mɑ̃.ʒe\
nous mangeâmes	\nu mɑ̃.ʒam\	nous eûmes mangé	\nu.z_ym mɑ̃.ʒe\
vous mangeâtes	\vu mɑ̃.ʒat\	vous eûtes mangé	\vu.z_yt mɑ̃.ʒe\
ils/elles mangèrent	\[il/ɛl] mɑ̃.ʒɛʁ\	ils/elles eurent mangé	\[i/ɛ]l.z_yʁ mɑ̃.ʒe\
Futur simple		**Futur antérieur**	
je mangerai	\ʒə mɑ̃.ʒ(ə.)ʁe\	j'aurai mangé	\ʒ_o.ʁe mɑ̃.ʒe\
tu mangeras	\ty mɑ̃.ʒ(ə.)ʁa\	tu auras mangé	\ty o.ʁa mɑ̃.ʒe\
il/elle/on mangera	\[il/ɛl/ɔ̃] mɑ̃.ʒ(ə.)ʁa\	il/elle/on aura mangé	\[i.l/ɛ.l/ɔ̃.n]_o.ʁa mɑ̃.ʒe\
nous mangerons	\nu mɑ̃.ʒ(ə.)ʁɔ̃\	nous aurons mangé	\nu.z_o.ʁɔ̃ mɑ̃.ʒe\
vous mangerez	\vu mɑ̃.ʒ(ə.)ʁe\	vous aurez mangé	\vu.z_o.ʁe mɑ̃.ʒe\
ils/elles mangeront	\[il/ɛl] mɑ̃.ʒ(ə.)ʁɔ̃\	ils/elles auront mangé	\[i/ɛ]l.z_o.ʁɔ̃ mɑ̃.ʒe\

Subjonctif

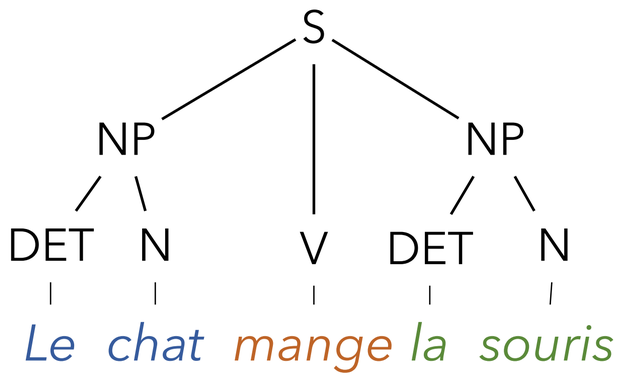
Présent		Passé	
que je mange	\kə ʒə mɑ̃ʒ\	que j'aie mangé	\kə ʒ_ɛ mɑ̃.ʒe\
que tu manges	\kə ty mɑ̃ʒ\	que tu aies mangé	\kə ty ɛ mɑ̃.ʒe\
qu'il/elle/on mange	\k_[il/ɛl/ɔ̃] mɑ̃ʒ\	qu'il/elle/on ait mangé	\k_[i.l/ɛ.l/ɔ̃.n]_ɛ mɑ̃.ʒe\
		que nous ayons mangé	\kə nu.z_ɛ.jɔ̃ mɑ̃.ʒe\
que nous mangions	\kə nu mɑ̃.ʒjɔ̃\	que vous ayez mangé	\kə vu.z_ɛ.je mɑ̃.ʒe\

Source: Wiktionary

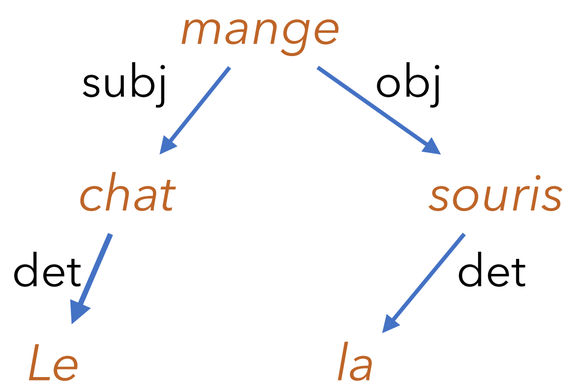
Syntax

How words are organised in grammatically correct sentences

- Constituents and dependencies are two ways to represent a sentence's syntactic structure
  - Constituents structure the sentence hierarchically (leaves in the tree are parts of speech)
  - Dependencies associate each word with their governor (the word they are attached to)



Constituency tree

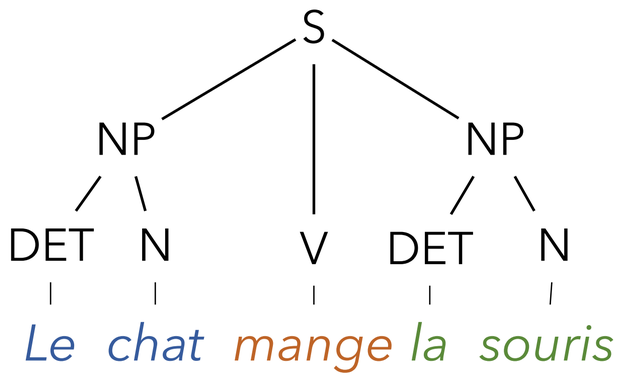


Dependency tree

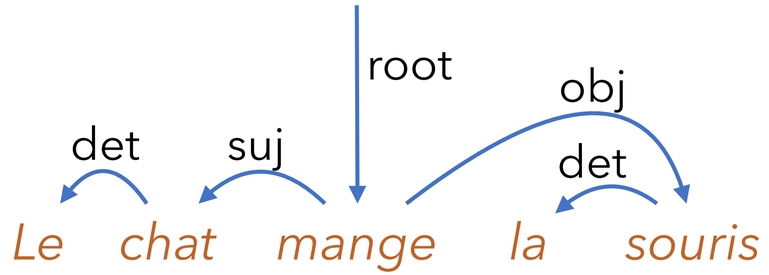
Syntax

How words are organised in grammatically correct sentences

- Constituents and dependencies are two ways to represent a sentence's syntactic structure
  - Constituents structure the sentence hierarchically (leaves in the tree are parts of speech)
  - Dependencies associate each word with their governor (the word they are attached to)



Constituency tree

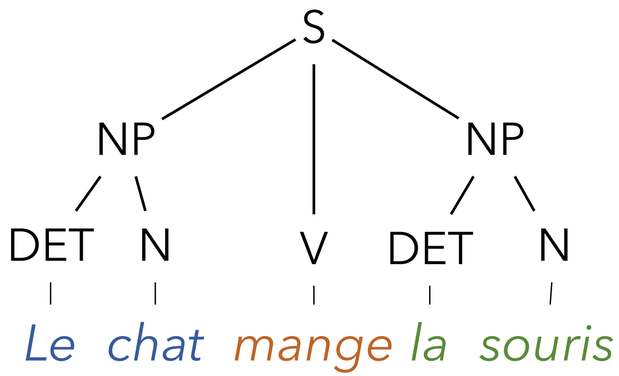


Dependency tree

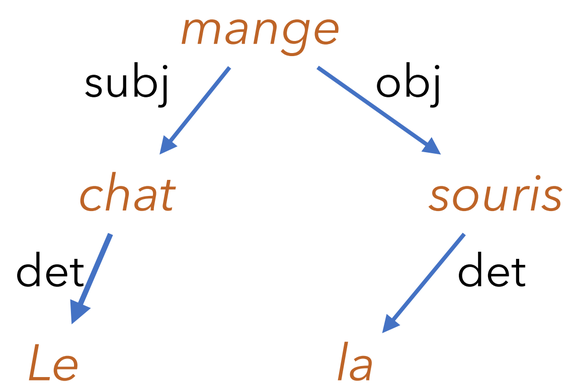
Syntax

How words are organised in grammatically correct sentences

- Constituents and dependencies are two ways to represent a sentence's syntactic structure
  - Constituents structure the sentence hierarchically (leaves in the tree are parts of speech)
  - Dependencies associate each word with their governor (the word they are attached to)



Constituency tree

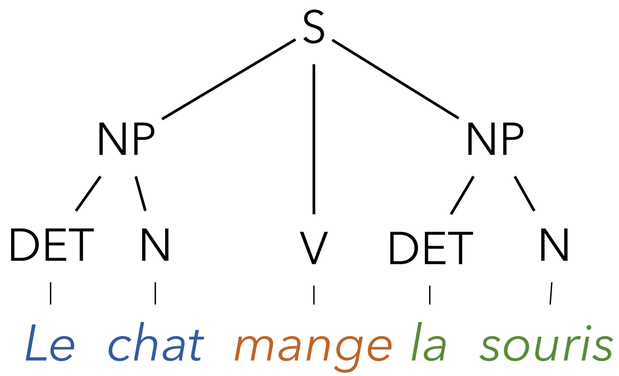


Dependency tree

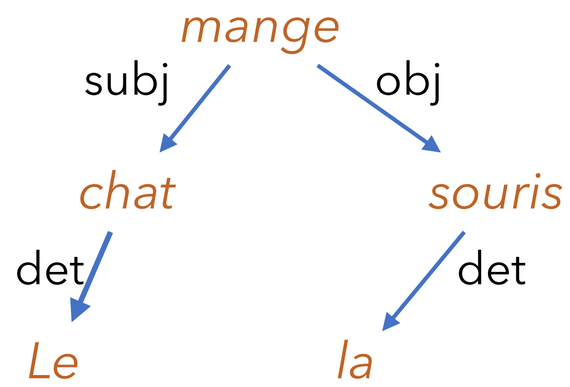
Syntax

How words are organised in grammatically correct sentences

- A dependency tree contains the information required to build the structure of the constituency tree
  - But information is missing to label internal nodes
- A constituency tree is not enough to recreate the dependency tree
  - Information is missing about the head of each constituent



Constituency tree

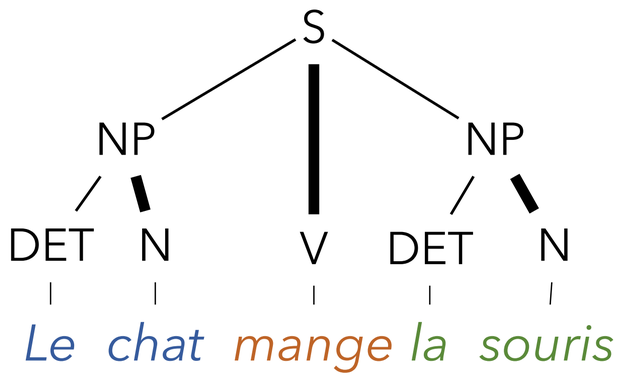


Dependency tree

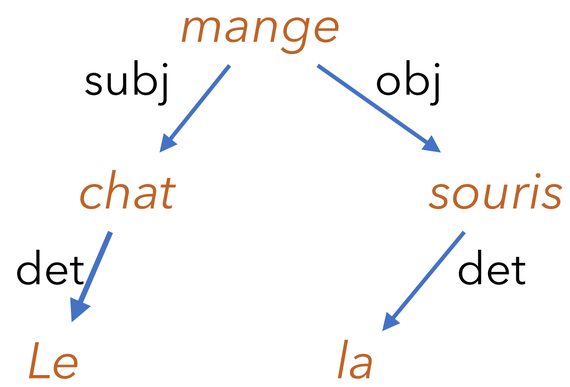
Syntax

How words are organised in grammatically correct sentences

- A dependency tree contains the information required to build the structure of the constituency tree
  - But information is missing to label internal nodes
- A constituency tree is not enough to recreate the dependency tree
  - Information is missing about the head of each constituent



Constituency tree



Dependency tree

Semantics

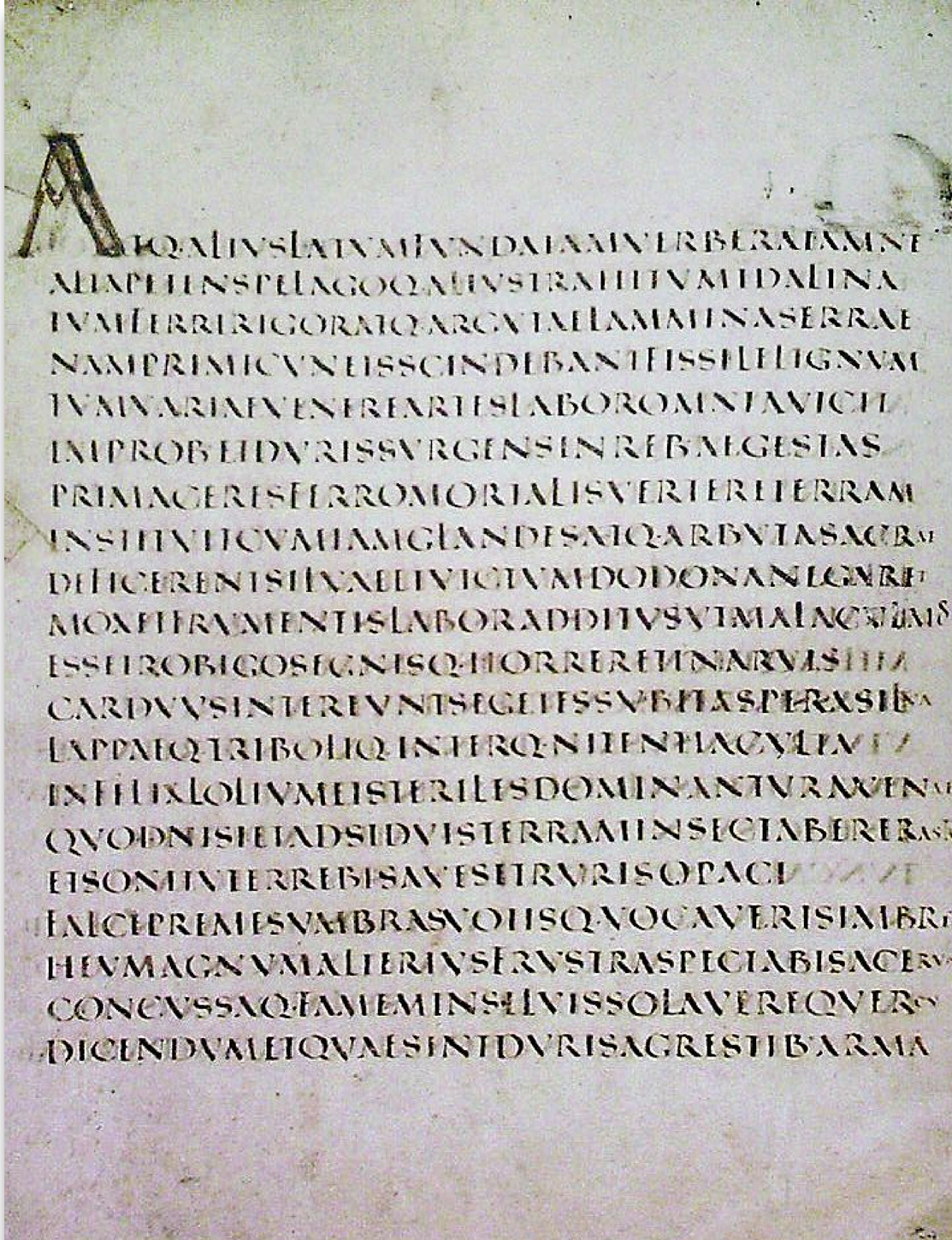
How words and sentences carry a meaning

- Lexical semantics: investigating the meaning of words and of idioms, exploring the relationships between semantically related words
- Formal semantics: logic-based modelling of how meaning is built from words and sentences
- Pragmatics: how context influences meaning
  - Linguistic context
  - Extra-linguistic context

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$. *DNI*$_{\text{TRIGGER}}$

Processing textual data

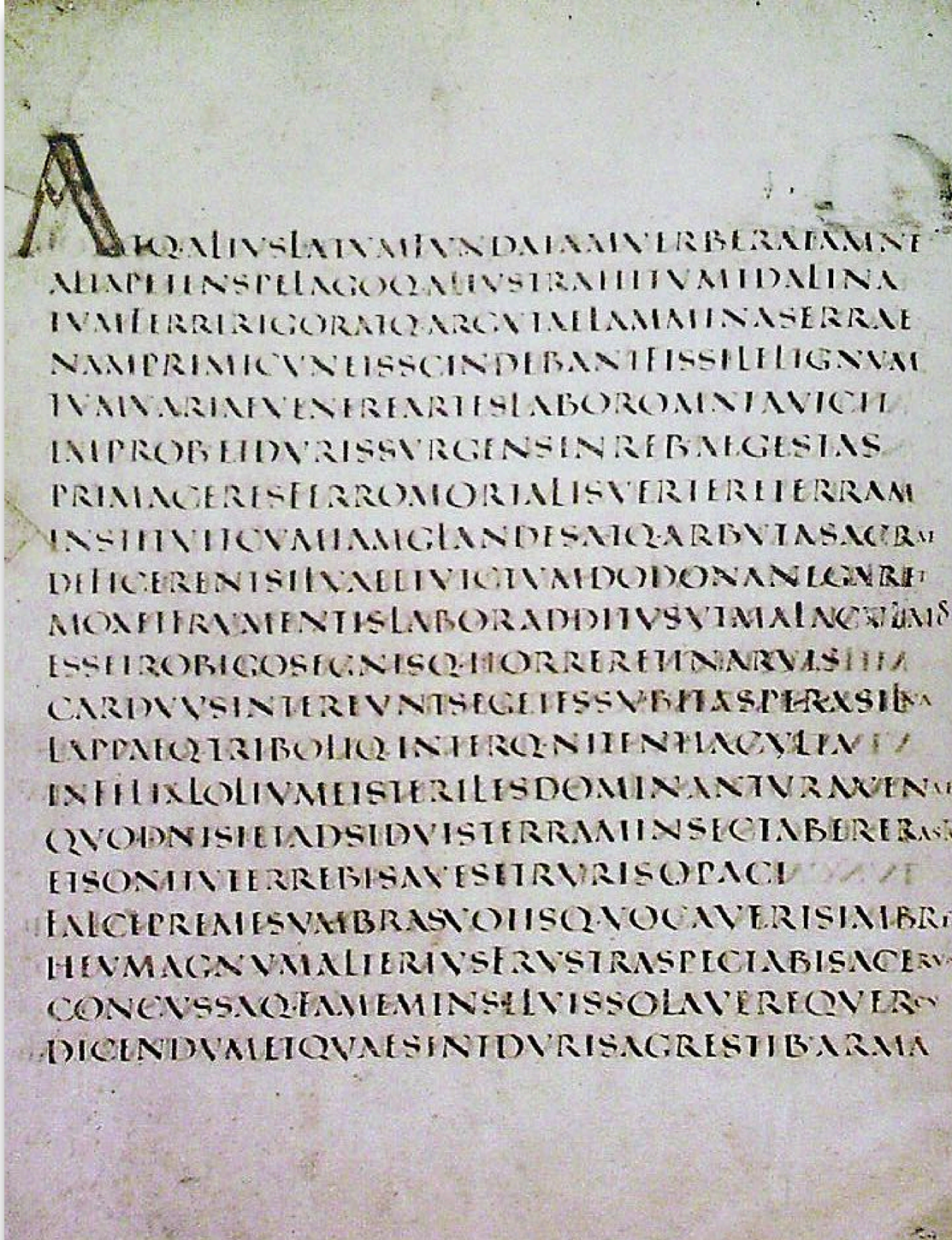
- A text is a **sequence of characters**
  - Letters, ideograms, syllabograms, and more
  - Punctuation marks
  - Separators (e.g. whitespace) – not in all writing systems
- NLP systems often rely on a **two-tier structuration** of sequences
  - Macroscopic units: **“sentences”**
  - Microscopic units: **“words”**
- Often, “sentences” are processed independently of one another, as sequences of “words”



Vergilius Augusteus, Georgica 141ff, exemple de *scriptio continua*. Source : Wikipedia

Processing textual data

- A text is a **sequence of characters**
  - Letters, ideograms, syllabograms, and more
  - Punctuation marks
  - Separators (e.g. whitespace) – not in all writing systems
- NLP systems often rely on a **two-tier structuration** of sequences
  - Macroscopic units: **“sentences”**
  - Microscopic units: **“words”**
- Often, “sentences” are processed independently of one another, as sequences of “words”
- This raises two questions:
  - ***How can we identify words and sentences in raw texts?***
  - ***How can we represent words?***



Vergilius Augusteus, Georgica 141ff, exemple de *scriptio continua*. Source : Wikipedia

What is a sentence?

- The output of a **macroscopic** segmentation process
  - **Complete syntactic structure**
    - Not that frequent in speech data
  - **Semantically related to other sentences**
  - **Typographically marked**
    - In the Latin script, the full stop is an ambiguous symbol, sometimes overloaded
    - Same thing for the capital letter at the beginning of a “sentence”
    - *Dès maintenant, la mobilisation est de mise. Pour l’amour des mots.*
      Best. Movie. Ever.
      L’épicerie vend des concombres, des salades, des radis, etc.
  - Embedded sentences, in particular with quotes:
    « Un dimanche matin, pour éviter le trafic routier, on a transporté la momie jusqu’à une clinique. Elle est conservée en permanence à – 20 °C pour que son état ne s’altère pas », se remémore-t-il. (Miren Garaicoechea, Le Monde, 3 décembre 2023)

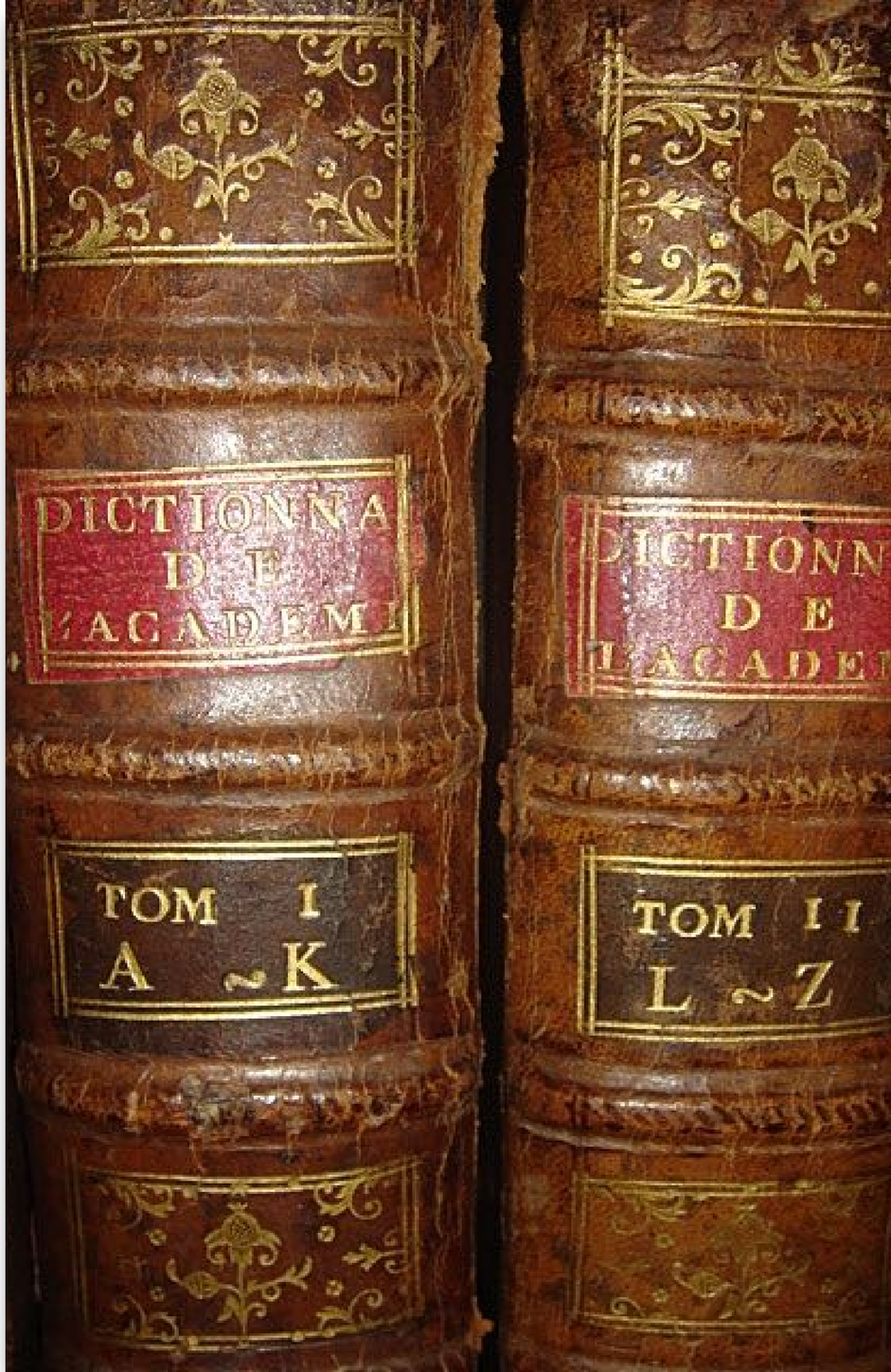
What is a sentence?

- The output of a **macroscopic** segmentation process
  - **Complete syntactic structure**
    - Not that frequent in speech data
  - **Semantically related to other sentences**
  - **Typographically marked**
    - In the Latin script, the full stop is an ambiguous symbol, sometimes overloaded
    - Same thing for the capital letter at the beginning of a “sentence”
    - *Dès maintenant, la mobilisation est de mise. Pour l’amour des mots.*
      Best. Movie. Ever.
      L’épicerie vend des concombres, des salades, des radis, etc.
  - Embedded sentences, in particular with quotes:
    « Un dimanche matin, pour éviter le trafic routier, on a transporté la momie jusqu’à une clinique. Elle est conservée en permanence à – 20 °C pour que son état ne s’altère pas », se remémore-t-il. (Miren Garaicoechea, Le Monde, 3 décembre 2023)

• Can you draw an illustration of what a sentence is, 16x9, vertically orientated? • try again • another one
DALL-E 3 (via ChatGPT, GPT-4, 06/12/2023)

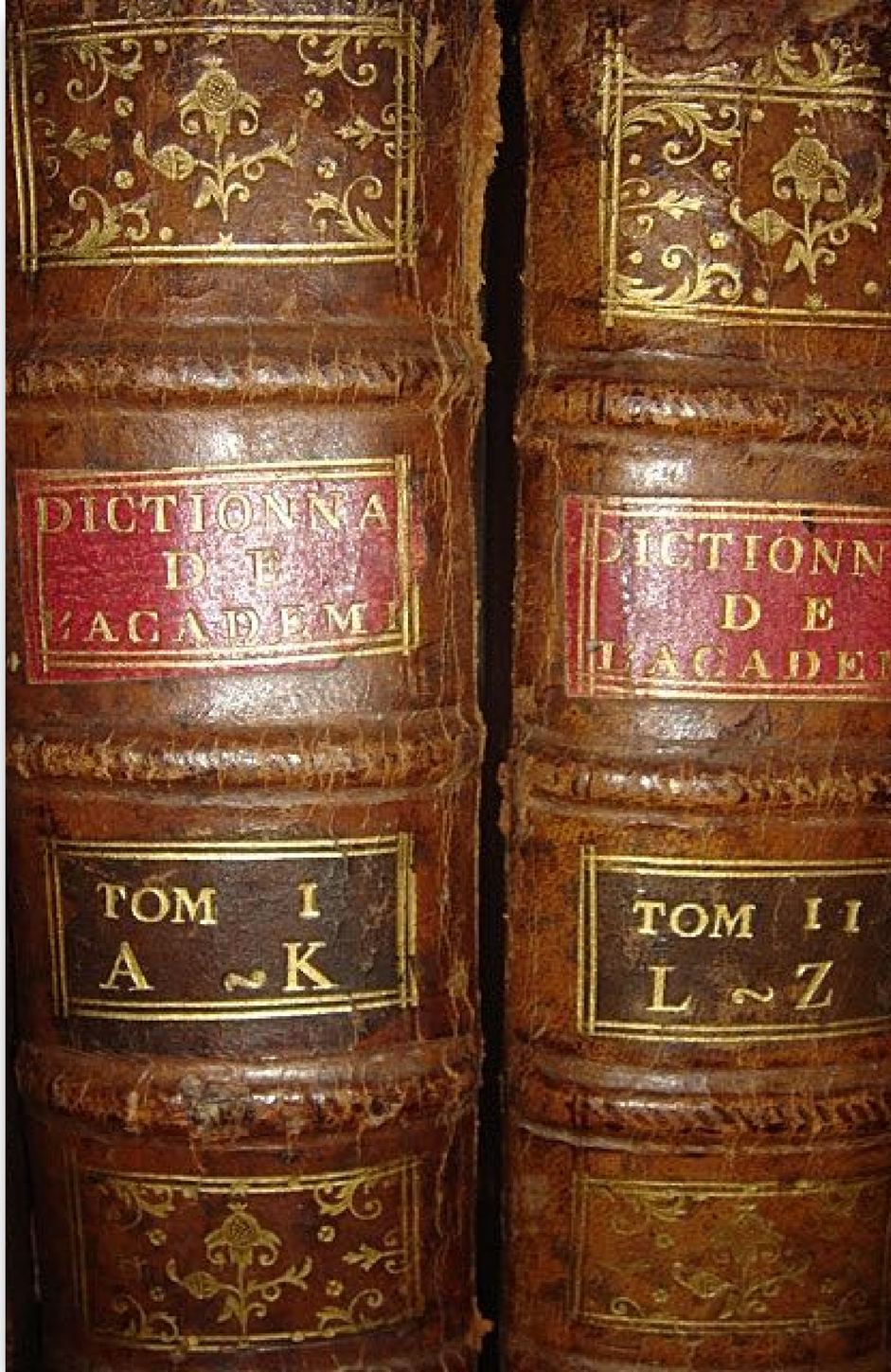
What is a word?

- No linguist would dare to suggest a unique definition for the notion of the “word”
- **At least four concepts must be distinguished**
  - The prosodic word
  - The typographic word, or token (or pre-token)
  - The morphosyntactic word, or word-form (or form)
  - The semantic word
- Mismatches between these concepts often reveal interesting linguistic questions
- These concepts approximately correspond to the **classical levels of linguistic analysis**



What is a word?

- No linguist would dare to suggest a unique definition for the notion of the “word”
- **At least four concepts must be distinguished**
  - The prosodic word
  - The typographic word, or token (or pre-token)
  - The morphosyntactic word, or word-form (or form)
  - The semantic word
- Mismatches between these concepts often reveal interesting linguistic questions
- These concepts approximately correspond to the **classical levels of linguistic analysis**



Source : Wikipedia (Heurtelions)

Pre-tokens (formerly “tokens”)

- A **pre-token** is a **purely typographic unit**, based on a non-ambiguous convention
- Starting point:
  - Multiple scripts use punctuation marks
  - Some use a typographic separator (such as the whitespace)
- In scripts that have such a separator, a pre-token can be defined as follows:
  - A character sequence containing neither separator nor punctuation mark,
  - or a punctuation mark

Tout à coup, il commença à jouer au tennis de table avec Pierre Martin.

Pre-tokens (formerly “tokens”)

- A **pre-token** is a **purely typographic unit**, based on a non-ambiguous convention
- Starting point:
  - Multiple scripts use punctuation marks
  - Some use a typographic separator (such as the whitespace)
- In scripts that have such a separator, a pre-token can be defined as follows:
  - A character sequence containing neither separator nor punctuation mark,
  - or a punctuation mark

Tout à coup , il commença à jouer au tennis de table avec Pierre Martin .

Pre-tokens (formerly “tokens”)

- A **pre-token** is a **purely typographic unit**, based on a non-ambiguous convention
- Starting point:
  - Multiple scripts use punctuation marks
  - Some use a typographic separator (such as the whitespace)
- In scripts that have such a separator, a pre-token can be defined as follows:
  - A character sequence containing neither separator nor punctuation mark,
  - or a punctuation mark

Tout à coup , il commença à jouer au tennis de table avec Pierre Martin .

Pre-tokens (formerly “tokens”)

- In scripts without a typographic separator, each character can be treated as a separate pre-token
- Examples: 我的漢語說得不太好。

  𒌷𒄩𒀜𒌅𒊭

Ḫattuša (𒌷𒄩𒀜𒌅𒊭), Porte des lions. Source : Wikipedia (C. Raddato)

Pre-tokens (formerly “tokens”)

- In scripts without a typographic separator, each character can be treated as a separate pre-token
- Examples: 我 的 漢 語 說 得 不 太 好 。

Ḫattuša (), Porte des lions. Source : Wikipedia (C. Raddato)

Tokens (or subwords)

- We sometimes need (especially in many widespread neural architectures) to use a pre-determined, fixed-size vocabulary (i.e. inventory of microscopic units)
- Idea: leave frequent “words” (tokens) unchanged and split less frequent ones in an “optimal” way
  - The units in the resulting vocabulary are called “**tokens**”, sometimes also “subwords”
  - They are obtained using algorithms and tools such as BPE (“byte pair encoding”), WordPiece, SentencePiece (see for instance Sennrich et al. (2016), Kudo & Richardson (2018))
- Subwords preserve the information regarding the presence of a separator

Tout à coup , il commença à jouer au tennis de table avec Pierre Martin .

Tokens (or subwords)

- We sometimes need (especially in many widespread neural architectures) to use a pre-determined, fixed-size vocabulary (i.e. inventory of microscopic units)
- Idea: leave frequent “words” (tokens) unchanged and split less frequent ones in an “optimal” way
  - The units in the resulting vocabulary are called “**tokens**”, sometimes also “subwords”
  - They are obtained using algorithms and tools such as BPE (“byte pair encoding”), WordPiece, SentencePiece (see for instance Sennrich et al. (2016), Kudo & Richardson (2018))
- Subwords preserve the information regarding the presence of a separator

Tout _à _coup _, _il commen ça _à _jou er _au _tennis _de _tab le _avec _Pierre _Martin _.

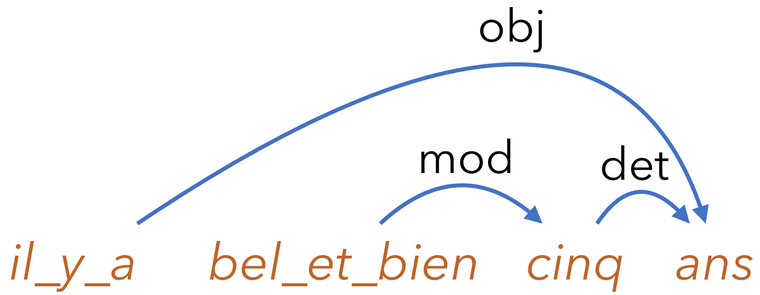
Tokens (or subwords)

- We sometimes need (especially in many widespread neural architectures) to use a pre-determined, fixed-size vocabulary (i.e. inventory of microscopic units)
- Idea: leave frequent “words” (tokens) unchanged and split less frequent ones in an “optimal” way
  - The units in the resulting vocabulary are called “**tokens**”, sometimes also “subwords”
  - They are obtained using algorithms and tools such as BPE (“byte pair encoding”), WordPiece, SentencePiece (see for instance Sennrich et al. (2016), Kudo & Richardson (2018))
- Subwords preserve the information regarding the presence of a separator

Tout _à _coup _, _il commen ça _à _jou er _au _tennis _de _tab le _avec _Pierre _Martin _.

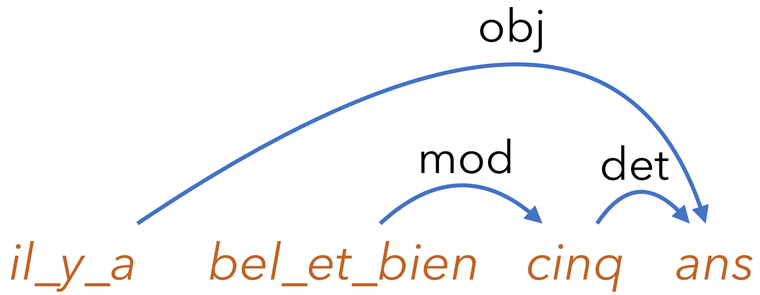
Word-forms (or forms)

- A word-form is a syntactically atomic unit: it is therefore a linguistically motivated unit
  - It can receive annotations (parts of speech: noun, adjective…; morphological features: plural, dative…) and corresponds to the leaves of a syntactic structure
- The pre-tokens⇔forms correspondence is neither simple nor deterministic
  - Structural mismatches take all possible forms:
    - *n* pre-tokens for 1 form = compound: *à l'instar de*, *bel et bien*, *il y a*
    - 1 pre-token for *m* forms = amalgams : *aux* = à + les
      - Sometimes ambiguous: *des* (1 token) = de + les (2 forms) ou des (1 form)
    - *n* pre-tokens for *m* forms: *à l' instar du* = à_l'_instar_de + le



Word-forms (or forms)

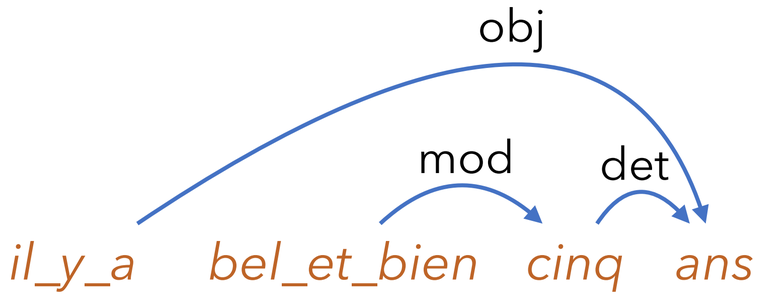
- A word-form is a syntactically atomic unit: it is therefore a linguistically motivated unit
  - It can receive annotations (parts of speech: noun, adjective…; morphological features: plural, dative…) and corresponds to the leaves of a syntactic structure
- The pre-tokens⇔forms correspondence is neither simple nor deterministic
  - Structural mismatches take all possible forms:
    - *n* pre-tokens for 1 form = compound: *à l'instar de*, *bel et bien*, *il y a*
    - 1 pre-token for *m* forms = amalgams : *aux* = à + les
      - Sometimes ambiguous: *des* (1 token) = de + les (2 forms) ou des (1 form)
    - *n* pre-tokens for *m* forms: *à l' instar du* = à_l'_instar_de + le



Tout à coup , il commença à jouer au tennis de table avec Pierre Martin .

Word-forms (or forms)

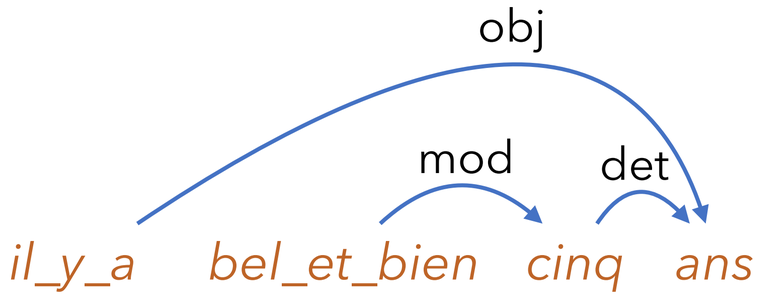
- A word-form is a syntactically atomic unit: it is therefore a linguistically motivated unit
  - It can receive annotations (parts of speech: noun, adjective…; morphological features: plural, dative…) and corresponds to the leaves of a syntactic structure
- The pre-tokens⇔forms correspondence is neither simple nor deterministic
  - Structural mismatches take all possible forms:
    - *n* pre-tokens for 1 form = compound: *à l'instar de*, *bel et bien*, *il y a*
    - 1 pre-token for *m* forms = amalgams : *aux* = à + les
      - Sometimes ambiguous: *des* (1 token) = de + les (2 forms) ou des (1 form)
    - *n* pre-tokens for *m* forms: *à l' instar du* = à_l'_instar_de + le



tout_à_coup , il commença à jouer à le tennis de table avec Pierre Martin .

Word-forms (or forms)

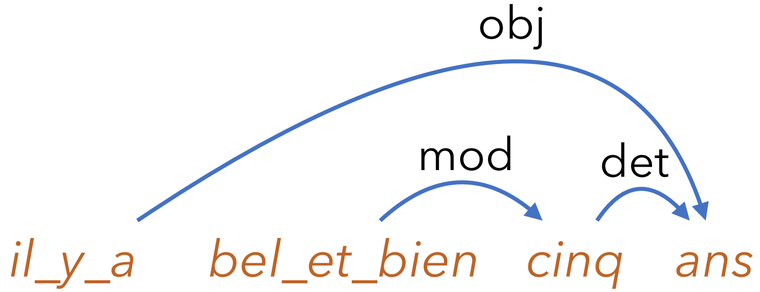
- A word-form is a syntactically atomic unit: it is therefore a linguistically motivated unit
  - It can receive annotations (parts of speech: noun, adjective…; morphological features: plural, dative…) and corresponds to the leaves of a syntactic structure
- The pre-tokens⇔forms correspondence is neither simple nor deterministic
  - Structural mismatches take all possible forms:
    - *n* pre-tokens for 1 form = compound: *à l'instar de*, *bel et bien*, *il y a*
    - 1 pre-token for *m* forms = amalgams : *aux* = à + les
      - Sometimes ambiguous: *des* (1 token) = de + les (2 forms) ou des (1 form)
    - *n* pre-tokens for *m* forms: *à l' instar du* = à_l'_instar_de + le



tout_à_coup , il commença à jouer à le tennis de table avec Pierre Martin .

Word-forms (or forms)

- A word-form is a syntactically atomic unit: it is therefore a linguistically motivated unit
  - It can receive annotations (parts of speech: noun, adjective…; morphological features: plural, dative…) and corresponds to the leaves of a syntactic structure
- The pre-tokens⇔forms correspondence is neither simple nor deterministic
  - Structural mismatches take all possible forms:
    - *n* pre-tokens for 1 form = compound: *à l'instar de*, *bel et bien*, *il y a*
    - 1 pre-token for *m* forms = amalgams : *aux* = à + les
      - Sometimes ambiguous: *des* (1 token) = de + les (2 forms) ou des (1 form)
    - *n* pre-tokens for *m* forms: *à l' instar du* = à_l'_instar_de + le



tout_à_coup , il commença à jouer à le tennis de table avec Pierre Martin .

Named entities

- A **named entity** is a real-world object that can be denoted individually
  - Standard named entities: people, locations, organisations
  - Extended named entities: dates, addresses, URLs, e-mail addresses, numbers, etc.
- A **named entity mention** is an utterance denoting a named entity
  - Examples: *Charles de Gaulle*, *Los Angeles*, *Apple Inc.*
  - This denotation can be ambiguous (*Orange*)
- Named entity mentions have specific properties:
  - Specific internal structure (**local grammar**), often more culture- than language-dependent ⇒ atomic for the general grammar ⇒ special word-forms

tout_à_coup , il commença à jouer à le tennis de table avec Pierre Martin .

Source: Wikipedia (JFK Library, Boston)

Named entities

- A **named entity** is a real-world object that can be denoted individually
  - Standard named entities: people, locations, organisations
  - Extended named entities: dates, addresses, URLs, e-mail addresses, numbers, etc.
- A **named entity mention** is an utterance denoting a named entity
  - Examples: *Charles de Gaulle*, *Los Angeles*, *Apple Inc.*
  - This denotation can be ambiguous (*Orange*)
- Named entity mentions have specific properties:
  - Specific internal structure (**local grammar**), often more culture- than language-dependent ⇒ atomic for the general grammar ⇒ special word-forms

tout_à_coup , il commença à jouer à le tennis de table avec Pierre_Martin .

Source: Wikipedia (JFK Library, Boston)

Named entities

- A **named entity** is a real-world object that can be denoted individually
  - Standard named entities: people, locations, organisations
  - Extended named entities: dates, addresses, URLs, e-mail addresses, numbers, etc.
- A **named entity mention** is an utterance denoting a named entity
  - Examples: *Charles de Gaulle*, *Los Angeles*, *Apple Inc.*
  - This denotation can be ambiguous (*Orange*)
- Named entity mentions have specific properties:
  - Specific internal structure (**local grammar**), often more culture- than language-dependent
    ⇒ atomic for the general grammar ⇒ special word-forms

tout_à_coup , il commença à jouer à le tennis de table avec Pierre_Martin .

Source: Wikipedia (JFK Library, Boston)

Semantic words

- A **semantic word** is a maximal sequence of forms whose meaning is **non-compositional**
  - Its meaning cannot be derived from that of its constitutive forms
  - Examples: *pomme de terre*, *red panda*, *peur bleue*, *red herring*

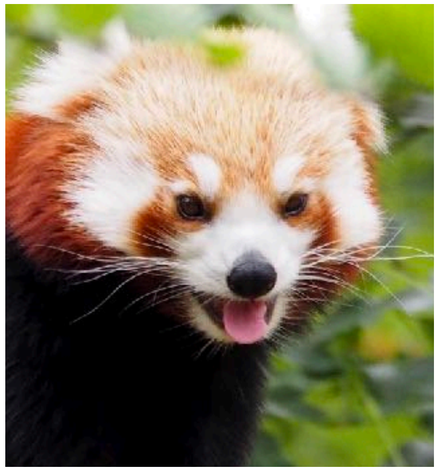


Photo BS

Semantic words

- A **semantic word** is a maximal sequence of forms whose meaning is **non-compositional**
  - Its meaning cannot be derived from that of its constitutive forms
  - Examples: *pomme de terre*, *red panda*, *peur bleue*, *red herring*
- Often ambiguous!
  Bob | *a* | *mangé* | *une* | *pomme de terre*
  Bob | *,* | *sculpteur* | *,* | *a* | *fabriqué* | *une* | *pomme* | *de* | *terre cuite*
- Close but distinct from the concept of **term** (conventional notion)
  Example: *machine à laver* (but not **appareil à nettoyer* !)

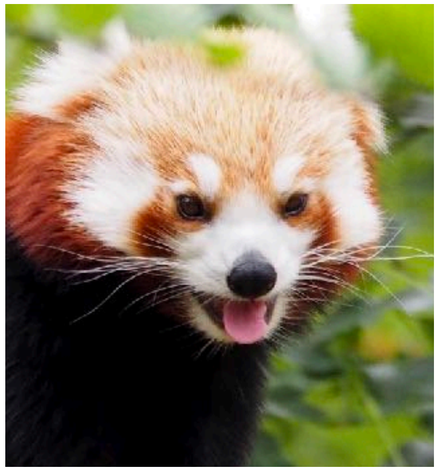


Photo BS

Source: Alibaba

Semantic words

- A **semantic word** is a maximal sequence of forms whose meaning is **non-compositional**
  - Its meaning cannot be derived from that of its constitutive forms
  - Examples: *pomme de terre*, *red panda*, *peur bleue*, *red herring*
- Often ambiguous!
  Bob | a | mangé | une | pomme de terre
  Bob | , | sculpteur | , | a | fabriqué | une | pomme | de | terre cuite
- Close but distinct from the concept of **term** (conventional notion)
  Example: *machine à laver* (but not **appareil à nettoyer* !)

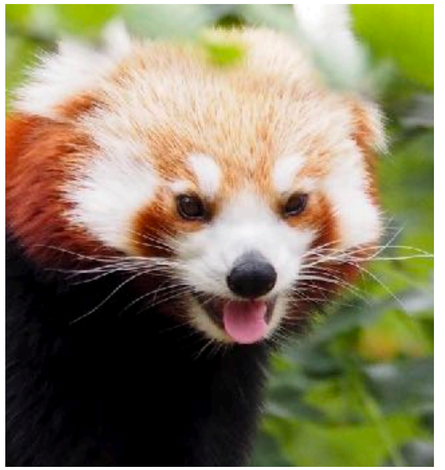


Photo BS

Source: Alibaba

tout_à_coup , il commença à jouer à le tennis de table avec Pierre_Martin .

Semantic words

- A **semantic word** is a maximal sequence of forms whose meaning is **non-compositional**
  - Its meaning cannot be derived from that of its constitutive forms
  - Examples: *pomme de terre*, *red panda*, *peur bleue*, *red herring*
- Often ambiguous!
  Bob | a | mangé | une | pomme de terre
  Bob | , | sculpteur | , | a | fabriqué | une | pomme | de | terre cuite
- Close but distinct from the concept of **term** (conventional notion)
  Example: *machine à laver* (but not **appareil à nettoyer* !)

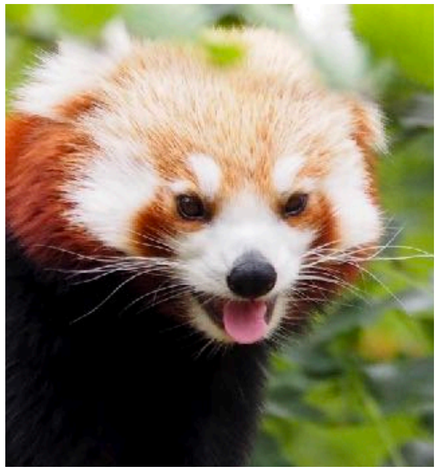


Photo BS

Source: Alibaba

tout_à_coup , il commença à jouer à le tennis_de_table avec Pierre_Martin .

4 specific challenges

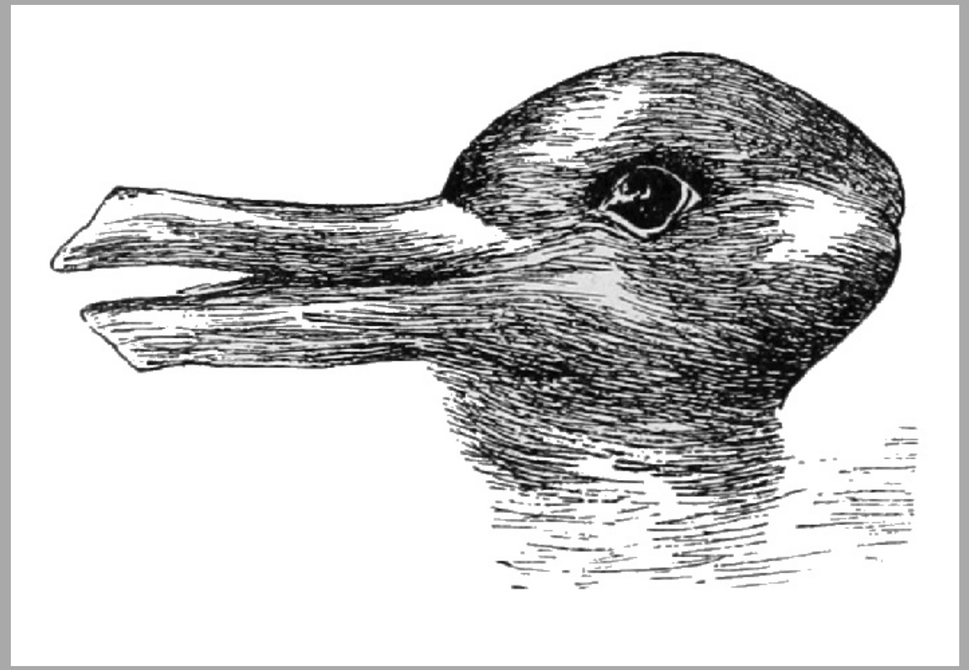
• *Draw textual data, steampunk style.*
DALL-E 3 (via ChatGPT, GPT-4, 06/12/2023)

4 specific challenges

- Language brings four fundamental challenges:
  - **Ambiguity** – meaning is often unclear
  - **Diversity** – there are many different languages working in different ways
  - **Variation** – content in a given language is diverse
  - **Sparsity** – many expressions are rare or unseen
- This makes (natural) languages very different from, for example, programming languages

• *Generate a vertical image illustrating the challenges of natural language: ambiguity, diversity, variation, sparsity.*
DALL-E 3 (via ChatGPT, GPT-5.2, 13/01/2026)

Ambiguity



Source: Wikipedia

Lexical ambiguity: homophony and homonymy

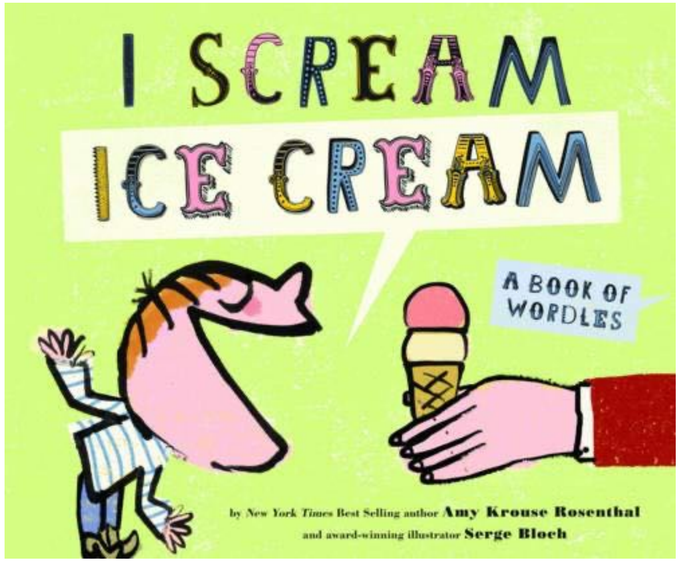
- Homophony: same pronunciation, different words (and often different spellings)
  - Examples in French: *vers*, *verre*, *ver*, *vert*, *vair*
- Extreme cases: holorhymes

  Étonnamment monotone et lasse

  Est ton âme en mon automne, hélas !

  (Louise de Vilmorin, 1954)
- Homography: same spelling, different words (and sometimes different pronunciations)
  - Example in French: *les poules du* ***couvent couvent***
  - Example in English: *if you have not* ***read*** *this book yet,* ***read*** *it!*

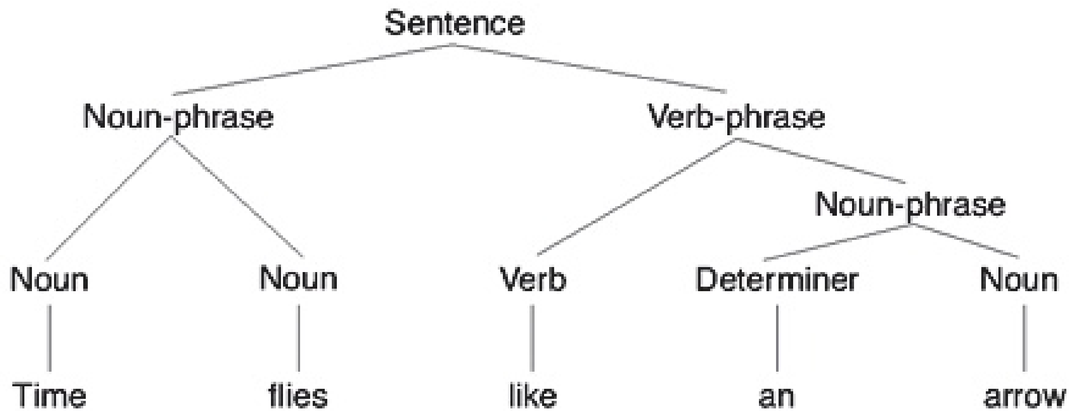
© Philippe Geluck



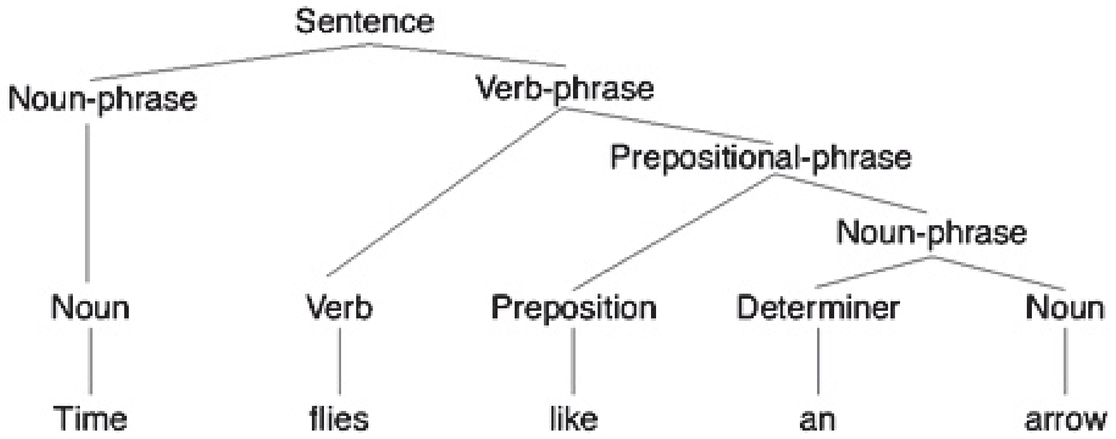
© Amy Krouse Rosenthal. I Scream, Ice Cream

Syntactic ambiguity

Time flies like an arrow



Fruit flies like a banana



Syntactic ambiguity: prepositional attachment

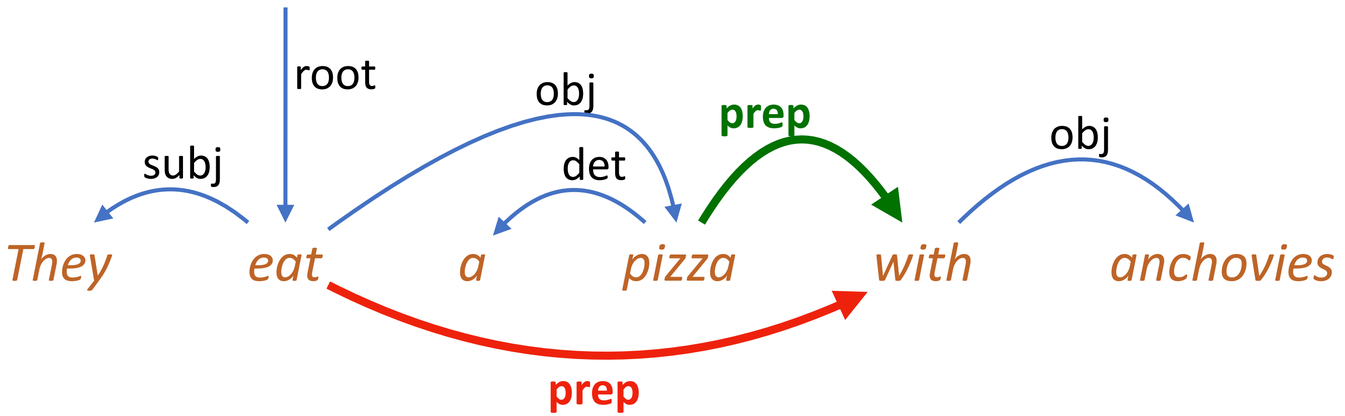


Illustration: CC-NC 2.5, James Constable (2010)

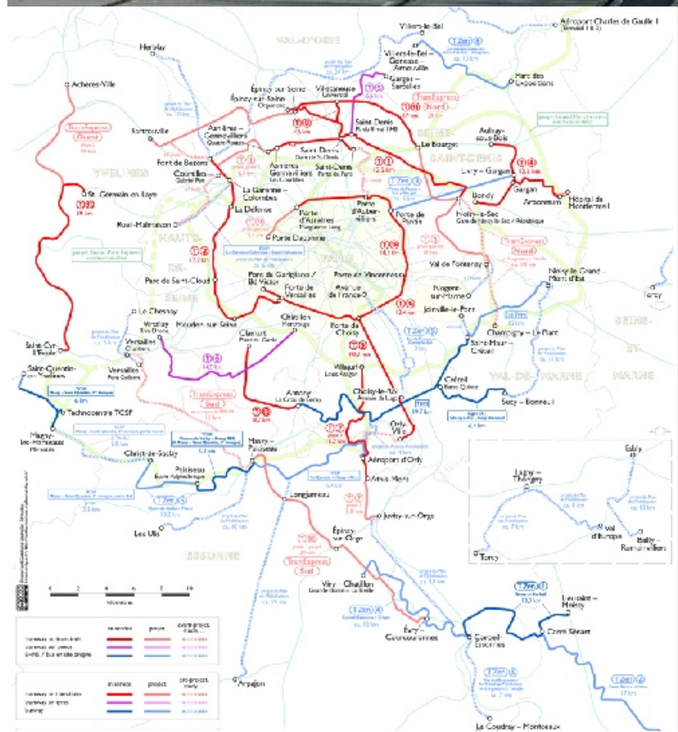
Intentional syntactic ambiguity

Photo BS (Las Vegas, 2013)

Semantic ambiguity

Polysemy

- Hyponymy/hyperonymy: *man* (vs. animal) ⊃ *man* (vs. woman) ⊃ *man* (vs. boy)
- Metaphor: *mole* (animal) > *mole* (spy)
- Object/colour: *orange* (fruit) > *orange* (colour)
- Object/informational content: *book* (object) // *book* (content)
- Object/collective abstract: *tramway* (vehicle) // *tramway* (means of transport)
- Tree or plant/what it produces: *cotton* (plant) > *cotton* (fibre/fabric)
- Animal/meat : *rabbit* (animal) > *rabbit* (meat)



Sources: Wikipedia

Diversity

Pieter Bruegel the Elder, *Tower of Babel*. Source: Wikipedia

Phonological diversity

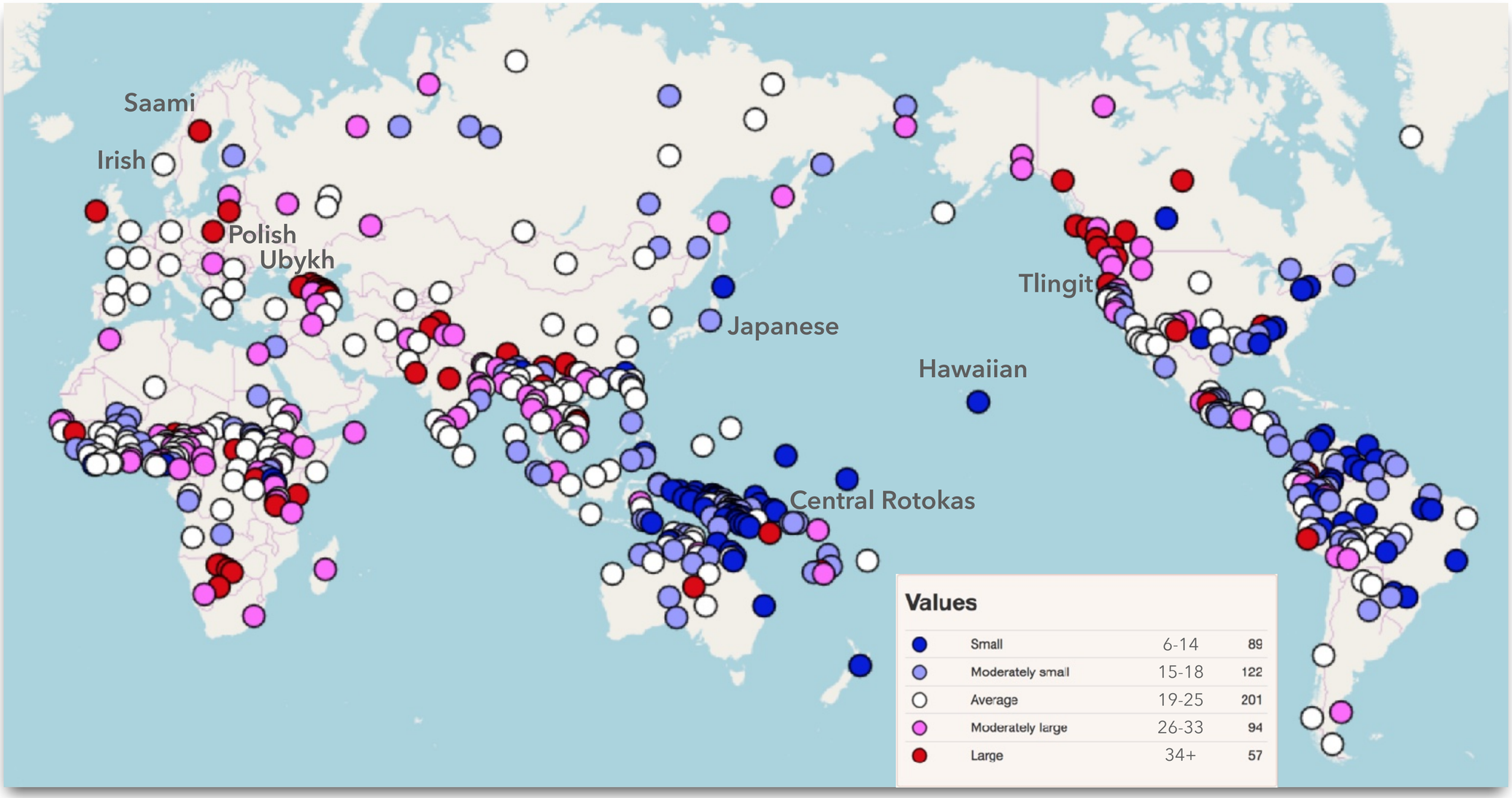


Figure: World Atlas of Language Structures - Consonant Inventories (Maddieson 2013)

Phonological diversity

Central Rotokas	Bilabial	Alveolar	Velar
Voiceless	p	t	k
Voiced	b ~ β	d ~ ɾ	g ~ ɣ

Ubykh		Labial		Alveolar			Postalveolar					Velar				Uvular					Glottal
							laminal closed		laminal		apical										
		plain	phar.	plain	lab.	lat.	plain	lab.	plain	lab.		pal.	plain	lab.	phar.	pal.	plain	lab.	phar.	phar. & lab.	
Plosive	voiceless	p	pˤ	t	tʷ							kʲ	k	kʷ		qʲ	q	qʷ	qˤ	qˤʷ	
	voiced	b	bˤ	d	dʷ							gʲ	g	gʷ							
	ejective	pʼ	pˤʼ	tʼ	tʷʼ							kʲʼ	kʼ	kʷʼ		qʲʼ	qʼ	qʷʼ	qˤʼ	qˤʷʼ	
Affricate	voiceless			t͡s			t͡ʃ		t͡ɕ	t͡ɕʷ	ʈ͡ʂ										
	voiced			d͡z			d͡ʒ		d͡ʑ	d͡ʑʷ	ɖ͡ʐ										
	ejective			t͡sʼ			t͡ʃʼ		t͡ɕʼ	t͡ɕʷʼ	ʈ͡ʂʼ										
Fricative	voiceless	f		s		ɬ	ʃ	ʃʷ	ɕ	ɕʷ	ʂ		x			χʲ	χ	χʷ	χˤ	χˤʷ	h
	voiced	v	vˤ	z			ʒ	ʒʷ	ʑ	ʑʷ	ʐ		ɣ			ʁʲ	ʁ	ʁʷ	ʁˤ	ʁˤʷ	
	ejective					ɬʼ															
Nasal		m	mˤ	n																	
Approximant						l						j		w	wˤ						
Trill				r																	

Source: Wikipedia

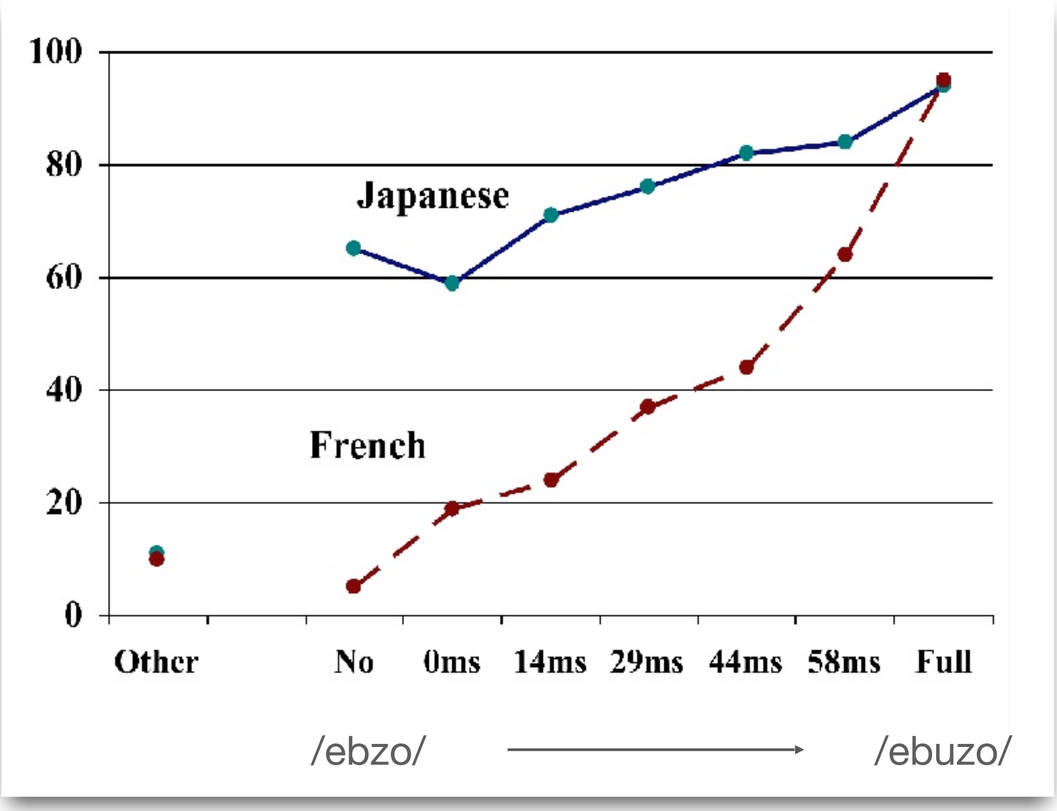
Phonological diversity

Syllables are formed of phoneme sequences

In most languages, some syllables are valid, some are not

Japanese: only V, CV, VN, CVN allowed

> phonological adaptation of borrowings:
sphinx > スフィンクス /sufiŋkusu/
Christmas > クリスマス /kurisumasu/



Phonological diversity

Syllables are formed of phoneme sequences

In most languages, some syllables are valid, some are not

Japanese: only V, CV, VN, CVN allowed

> phonological adaptation of borrowings:
sphinx > スフィンクス /sufiŋkusu/
Christmas > クリスマス /kurisumasu/

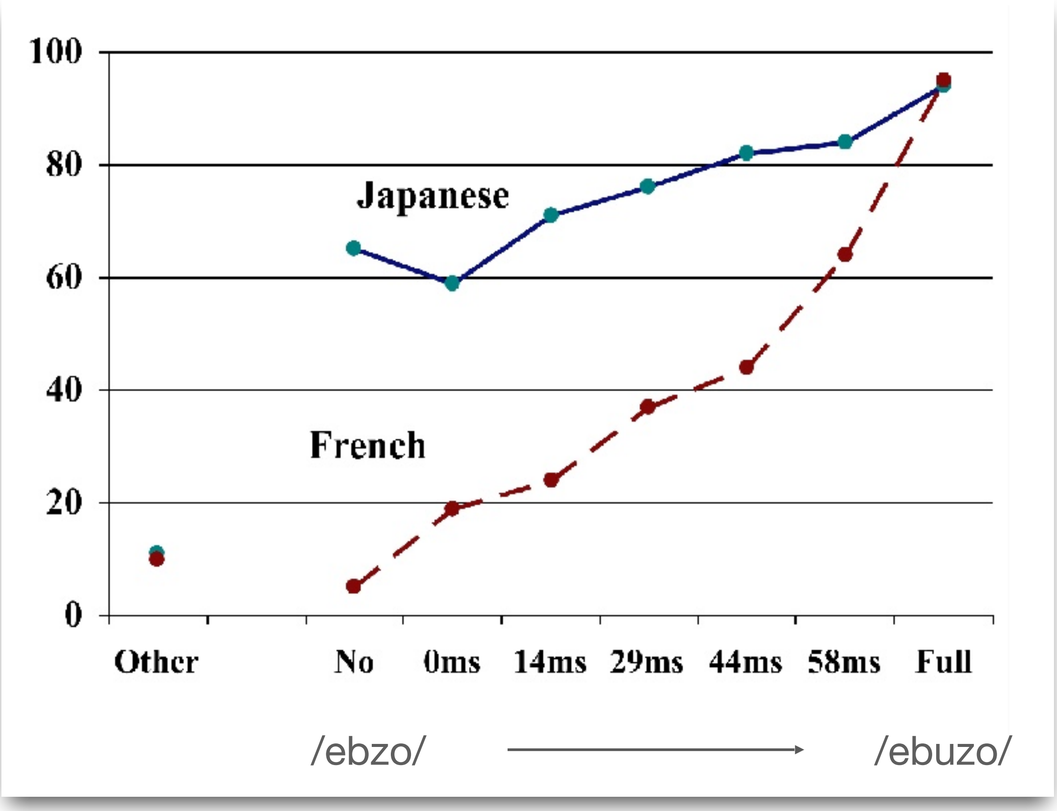


Figure adapted from (Dupoux *et al.* 1998)

Phonological diversity

Different vowel/consonant frequencies and cluster usage:

- Czech/Slovak *strč prst skrz krk* 'stick a finger through the neck'
- Nuxalk (Bella Coola) *clhp'xwlhtlhplhhskwts'* /xɬpʼχʷɬtʰɬpʰɬːskʷʰt͡sʼ/ 'he had possessed a bunchberry plant'
- Hawaiian *He aha kēia?* 'What is it?'

Source: Wikipedia (JohnHarvey, Peter coxhead)

Morphological diversity

- **Isolating languages**: each word carries exactly one meaning
  - E.g. Mandarin /ʊ̯ɔ^{214} mən^{4} tʰa̲n^{35} kɑŋ^{55}t͡ɕʰin^{35} lə^{5}/ (1p plur PLAY PIANO past) ‘we played the piano’
- **Synthetic languages**
  - **Agglutinative**: each word can have several morphs, each carrying one meaning
    - E.g. Turkish *el-ler-imiz-in* (HAND-pl-poss1pl-genitive) ‘of our hands’
  - **Fusional**: each word can have several morphs, each carrying one or more meanings, of which only one lexical morph
    - E.g. Latin *rexistis* /rek-s-is-tis/ (RULE-perf-perf-perf.2sg) ‘you$_{PLUR}$ ruled’
  - **Polysynthetic**: each word can have several lexical or grammatical morphs
    - E.g. Tuscarora (Iroquoien) *waʔ-khe-ta ʔnar-atya ʔt-hahθ* (passé-1sS/3fO-BREAD-BUY-appl:punc) ‘I bought bread for her’ (Williams 1976)

Morphological diversity

- Most languages show elements of different morphological types
- Including English and French
  - *le chien va jouer avec le chat*
  - *John’s cat eats mice*
  - *irions*
- Other example: creating words from whole sentences
  - French: *je-m’en-foutisme*
  - English: *You know, I can't take all this let's-be-faithful-and-never-look-at-another-person routine, because it just doesn't work* (The Boys in the Band, 1970)

Syntactic diversity

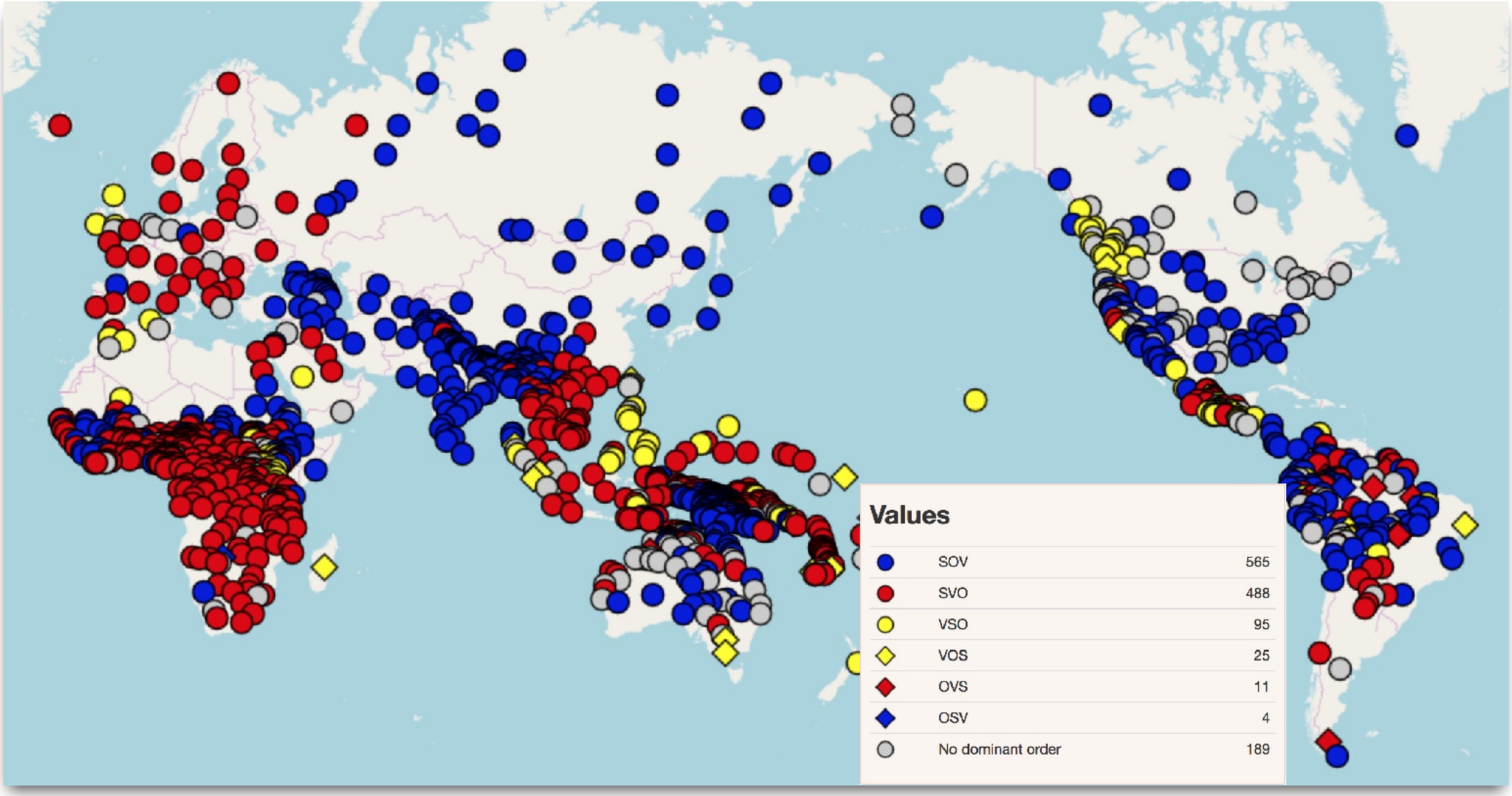


Figure: World Atlas of Language Structures - Order of S, O and V (Dryer 2013)

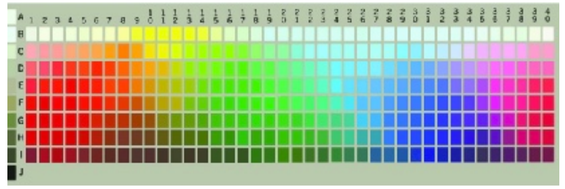
Syntactic diversity: levels of configurationality

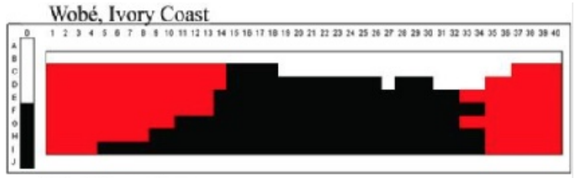
- Languages with **free word order**, often with very rich morphological marking and frequent discontinuous constituents
  - Example: Warlpiri (Australia)
- Languages with **partially free word order**, often with rich morphological marking and some discontinuous constituents
  - Example: Polish ‘Jean went to the cinema’ *Jaś poszedł do kina.*
    Poszedł Jaś do kina.
    Jaś do kina poszedł.
    Poszedł do kina Jaś.
    Do kina Jaś poszedł.
    Do kina poszedł Jaś.
- Languages with **fixed word order** (“**configurational**”), often with limited morphological marking and rare discontinuous constituents, if any
  - Examples: English, Mandarin Chinese

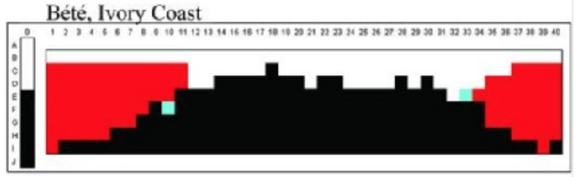
Semantic diversity

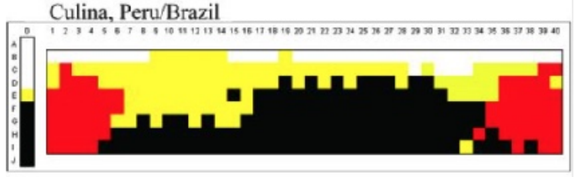
Words (fuzzily) partition the semantic space

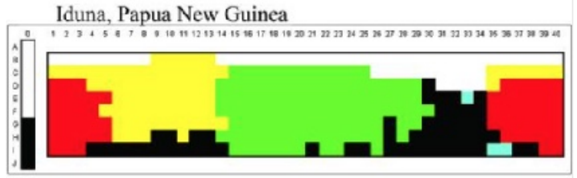
Partitions can differ from one language to another

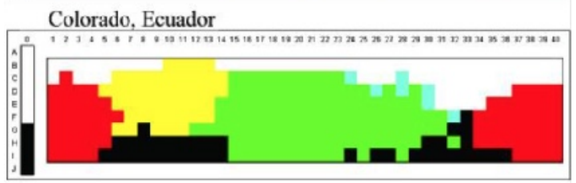


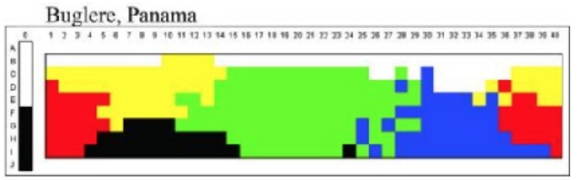


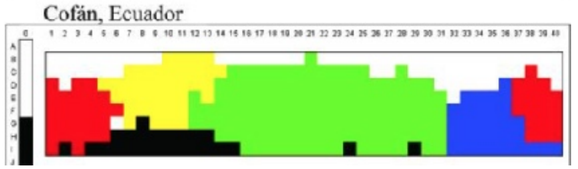


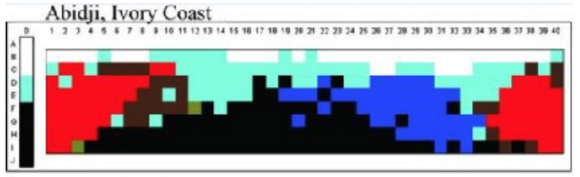


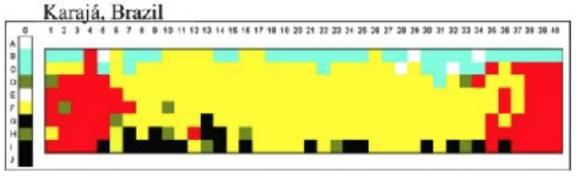


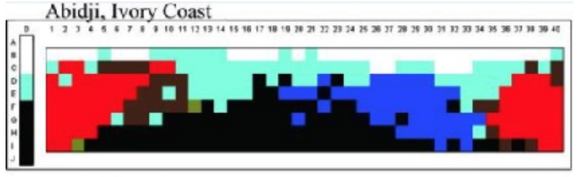


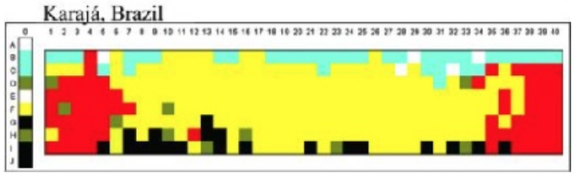










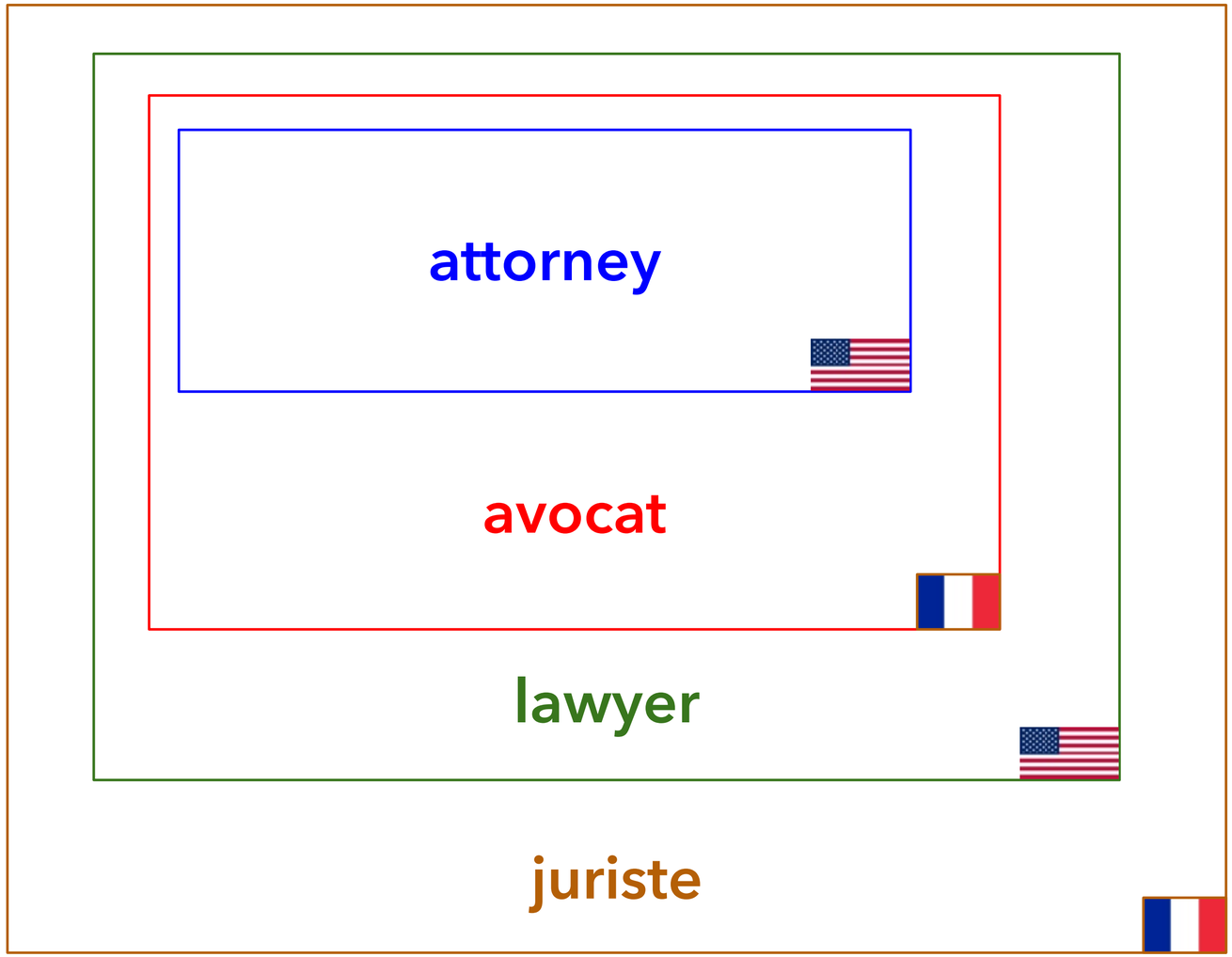


Source: World Colour Survey (Kay et al. 2009)

Lexical-cultural diversity

Words (fuzzily) partition the semantic space

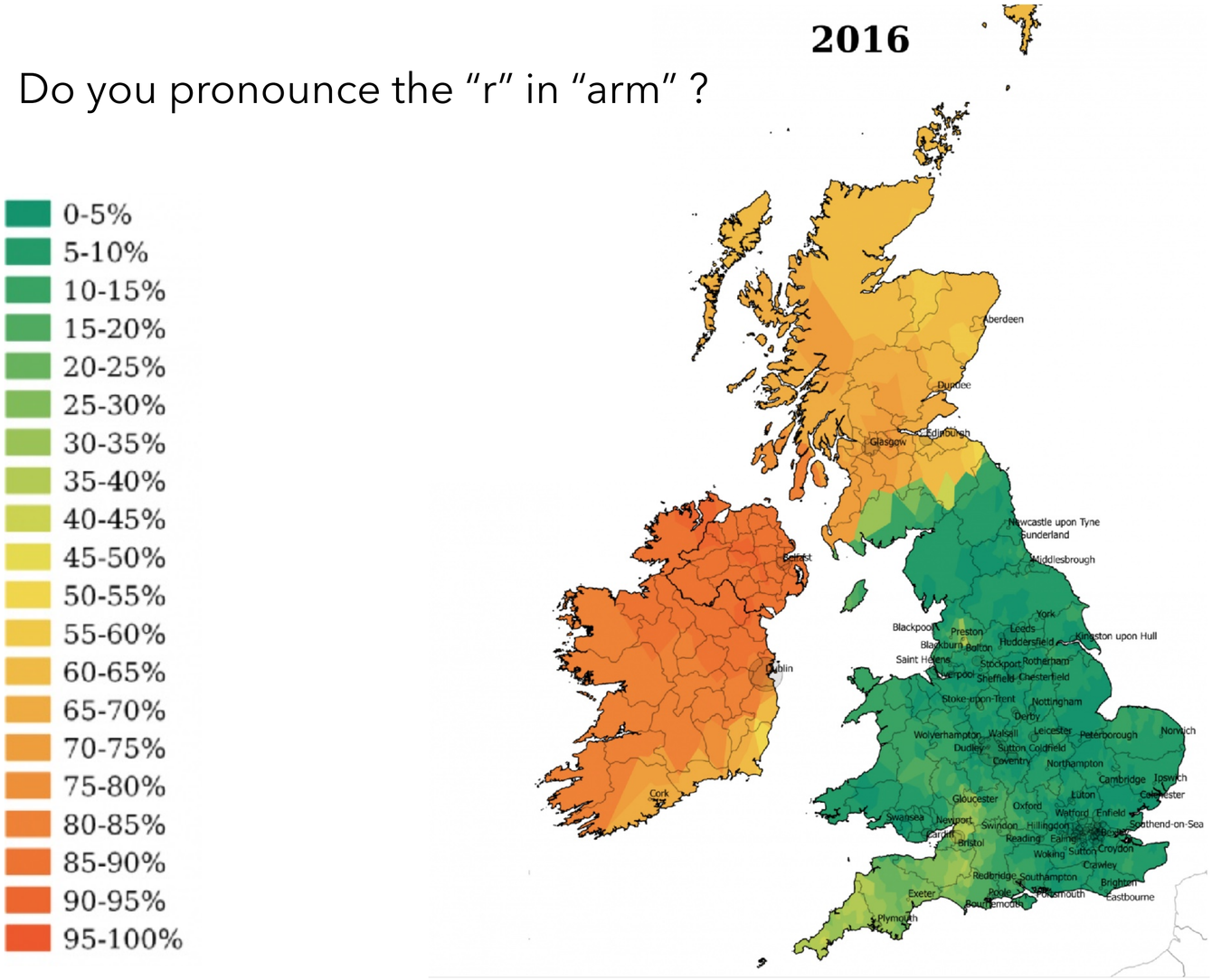
Partitions can differ from one language to another



Variation

Phonetic and phonological variation

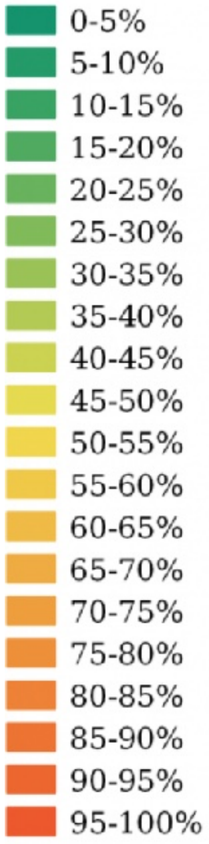
Do you pronounce the “r” in “arm” ?

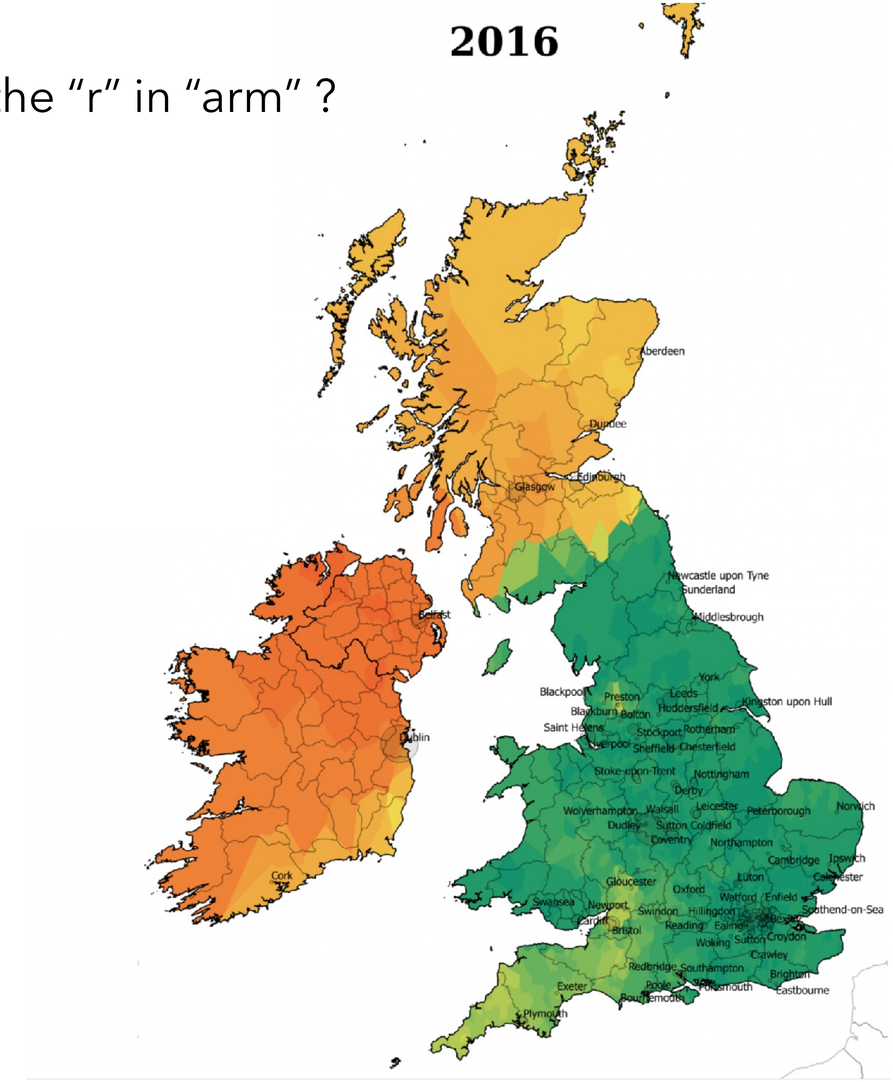


Source: University of Cambridge (2016)

Phonetic and phonological variation

Do you pronounce the “r” in “arm” ?





Screenshot of *The Voice Recognition Lift - ELEVEN!*
(with Iain Connell and Robert Florence)

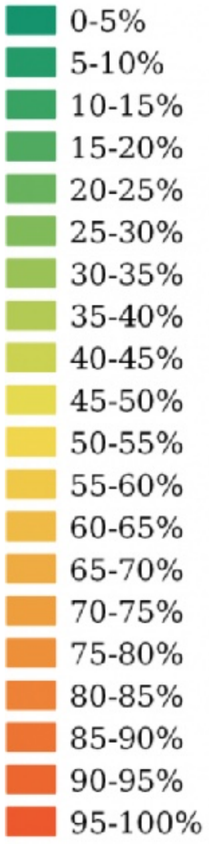
Source: University of Cambridge (2016)

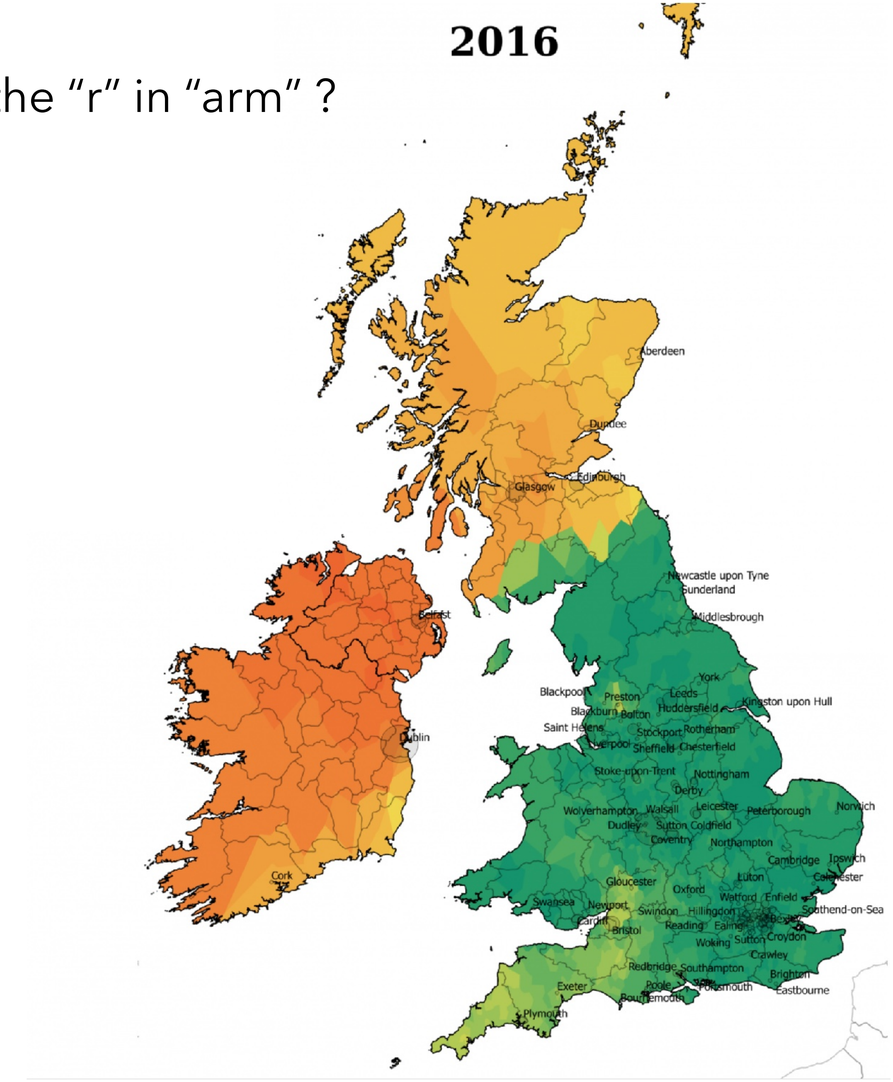
Phonetic and phonological variation

Phonetic and phonological variation

Phonetic and phonological variation

Do you pronounce the “r” in “arm” ?



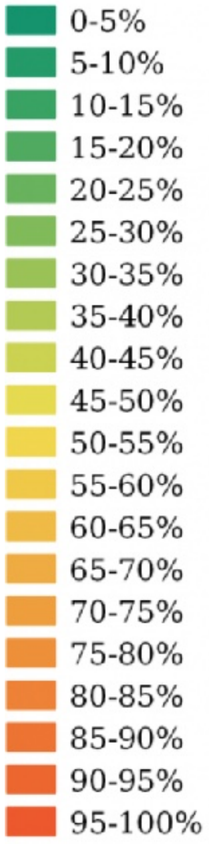


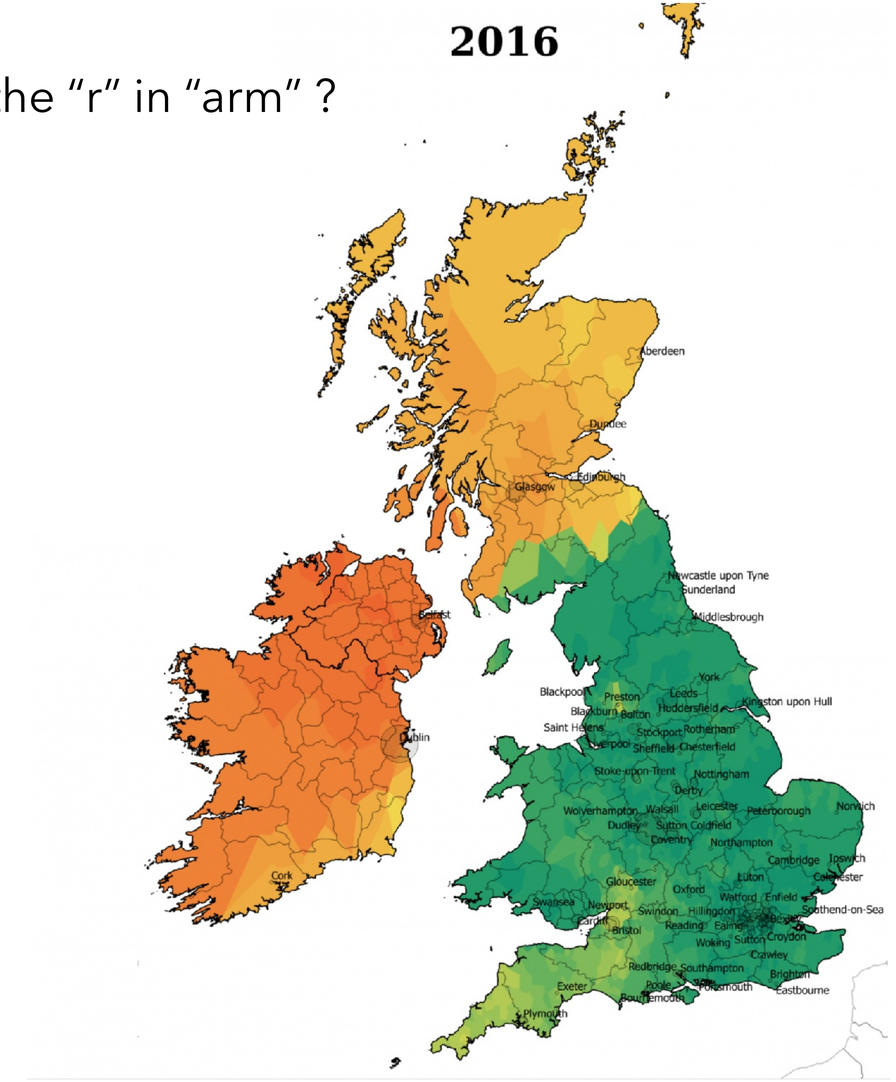
Screenshot of *The Voice Recognition Lift - ELEVEN!*
(with Iain Connell and Robert Florence)

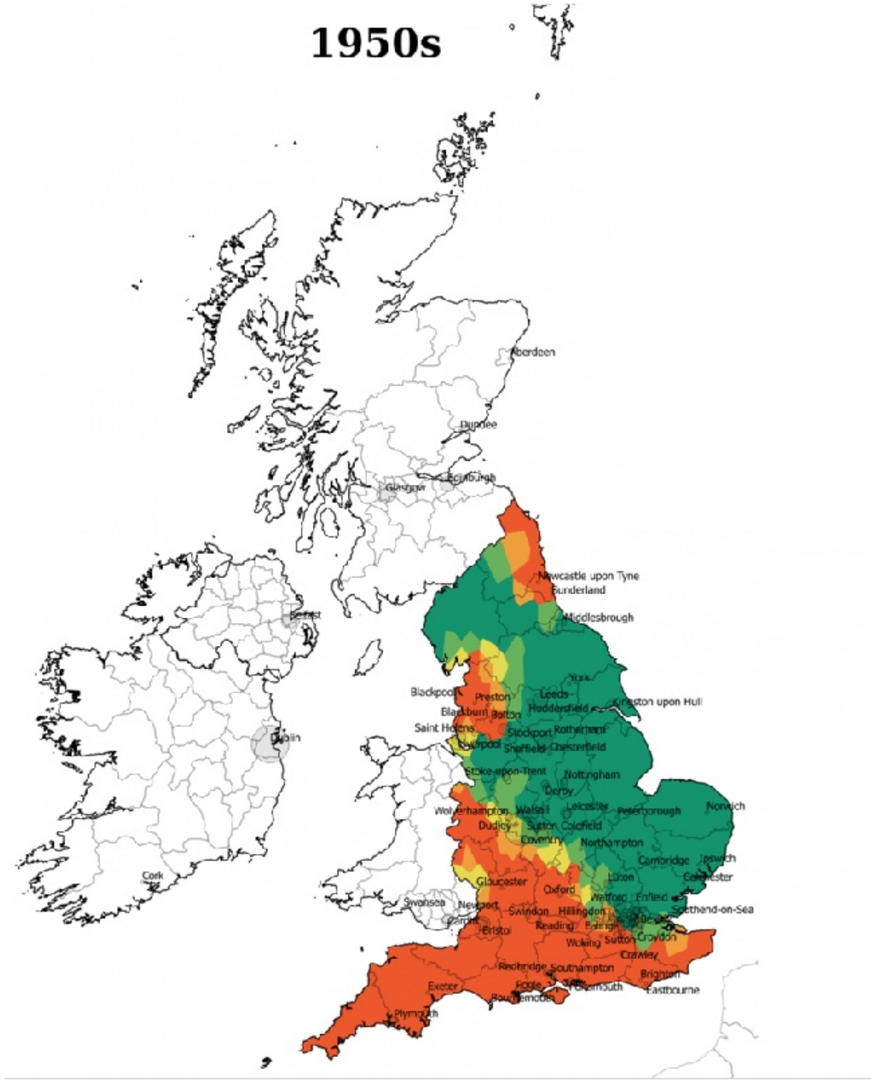
Source: University of Cambridge (2016)

Phonetic and phonological variation

Do you pronounce the “r” in “arm” ?







Source: University of Cambridge (2016)

Spelling variation

anagement maagement maanagement

maangement magagement magement

mamagement mamangement manaagement manaement

managaement manageement manageemnt managegment

managemaent managemant managememt managemen managemenet

managementt managemet managemetn managemnent managemnet

managemnt managemrnt managemt managenent managenment managent

managerment managhement managmeent managmement managment managnment

manament manamgement mananement manangment manasgement

manegement manegment mangaement mangagement mangagment

mangament mangement manggement mangment

mangmt menagement mgmt mgnt

mnagement mngmnt mngmt

Sociolinguistic variation

T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> tabuz, lavé rien fé
> ki ca ? le mec ou son chien ?
> Wtf is wrong with him ? #PETA4EVER
> ki ca ? le chien ?
> looool

Photo: Nikolas Giakoumidis/AP
Text: Twitter @rigolboche, inspired by other tweets
Translation: Bing Translation (06/12/2023)

Sociolinguistic variation

T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> tabuz, lavé rien fé
> ki ca ? le mec ou son chien ?
> Wtf is wrong with him ? #PETA4EVER
> ki ca ? le chien ?
> looool

BING translation :
Did you see he looked for it wsh#AperoChezRicard
> +10000, shah!
> Tabuz, washed nothing fe
> What is it? The guy or his dog?
> Wtf is wrong with him? #PETA4EVER
> What is it? The dog?
> looool

Photo: Nikolas Giakoumidis/AP
Text: Twitter @rigolboche, inspired by other tweets
Translation: Bing Translation (06/12/2023)

Diachronic variation

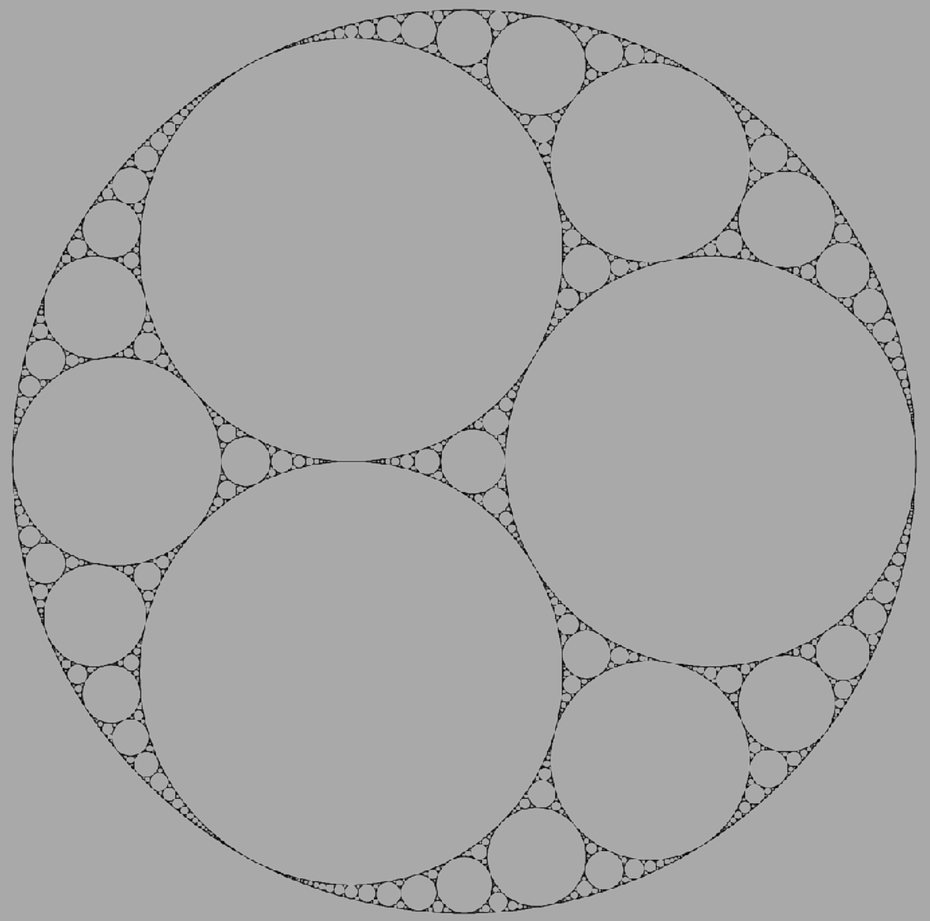
Li reis Marsilie esteit en Sarraguce.
Alez en est en un verger suz l'umbre;
Sur un perrun de marbre bloi se culchet,
Envirun lui plus de vint milie humes.
Il en apelet e ses dux e ses cuntes:
« Oëz, seignurs, quel pecchet nus encumbret :
Li emper[er]es Carles de France dulce
En cest païs nos est venuz cunfundre.
Jo nen ai ost qui bataille li dunne,
Ne n'ai tel gent ki la sue derumpet.
Cunseilez mei cume mi savie hume,
Si m(e) guarisez e de mort et de hunte. »
N'i ad paien ki un sul mot respundet,
Fors Blancandrins de Castel de Valfunde.

Chanson de Roland (11th century)

Hwæt! Wé Gárdena in géardagum
þéodcyninga þrym gefrúnon·
hú ðá æþelingas ellen fremedon.
Oft Scyld Scéfing sceaþena þréatum
monegum maégþum meodosetla oftéah·
egsode Eorle syððan aérest wearð
féasceaft funden hé þæs frófre gebád·
wéox under wolcnum· weorðmyndum þáh
oð þæt him aéghwylc þára ymbsittendra
ofer hronráde hýran scolde,
gomban gyldan· þæt wæs gód cyning.

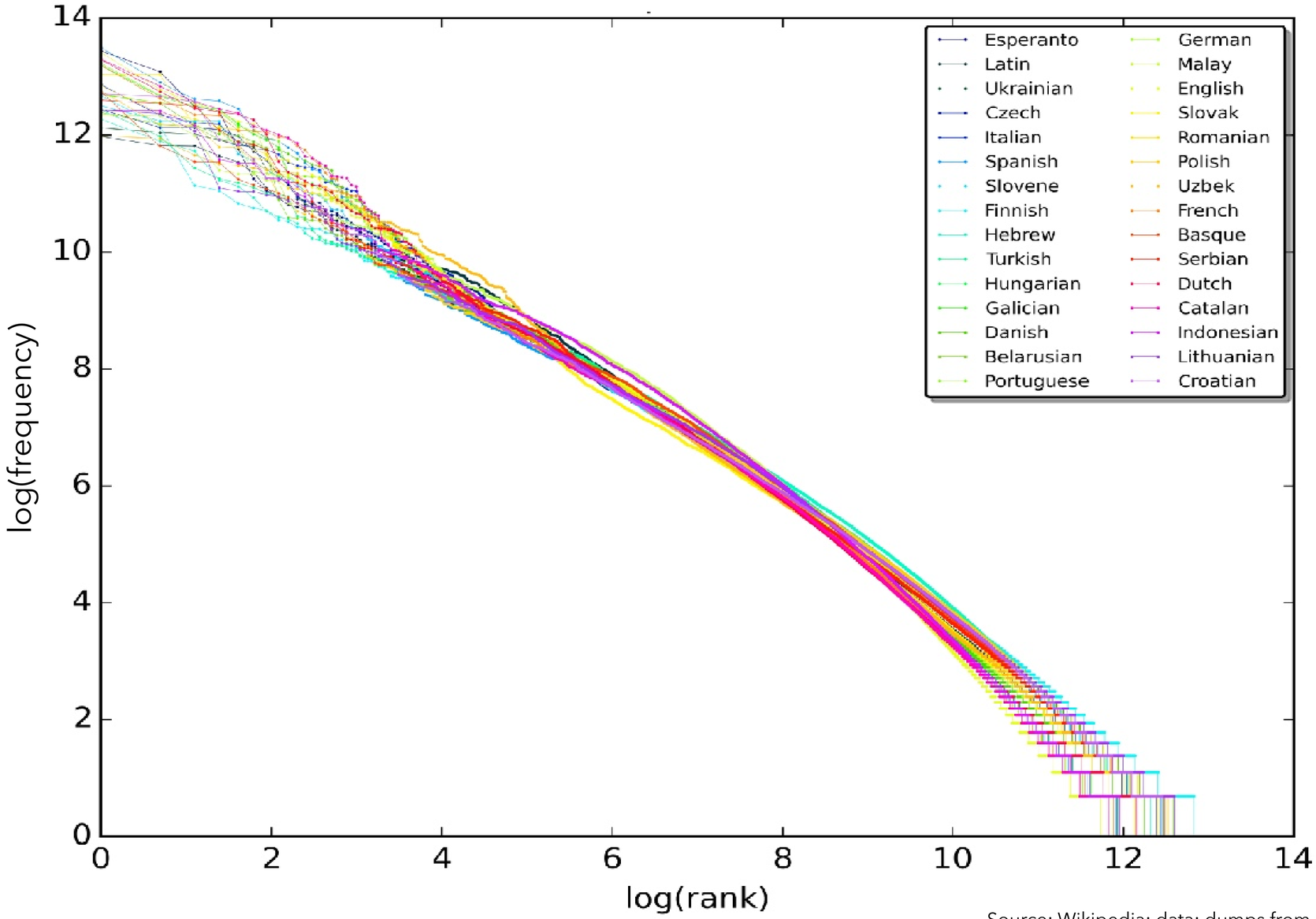
Beowulf (10th century)

Sparsity



Zipf's law

Graph of Rank / frequency for the first 10 million words in 30 wikipedias

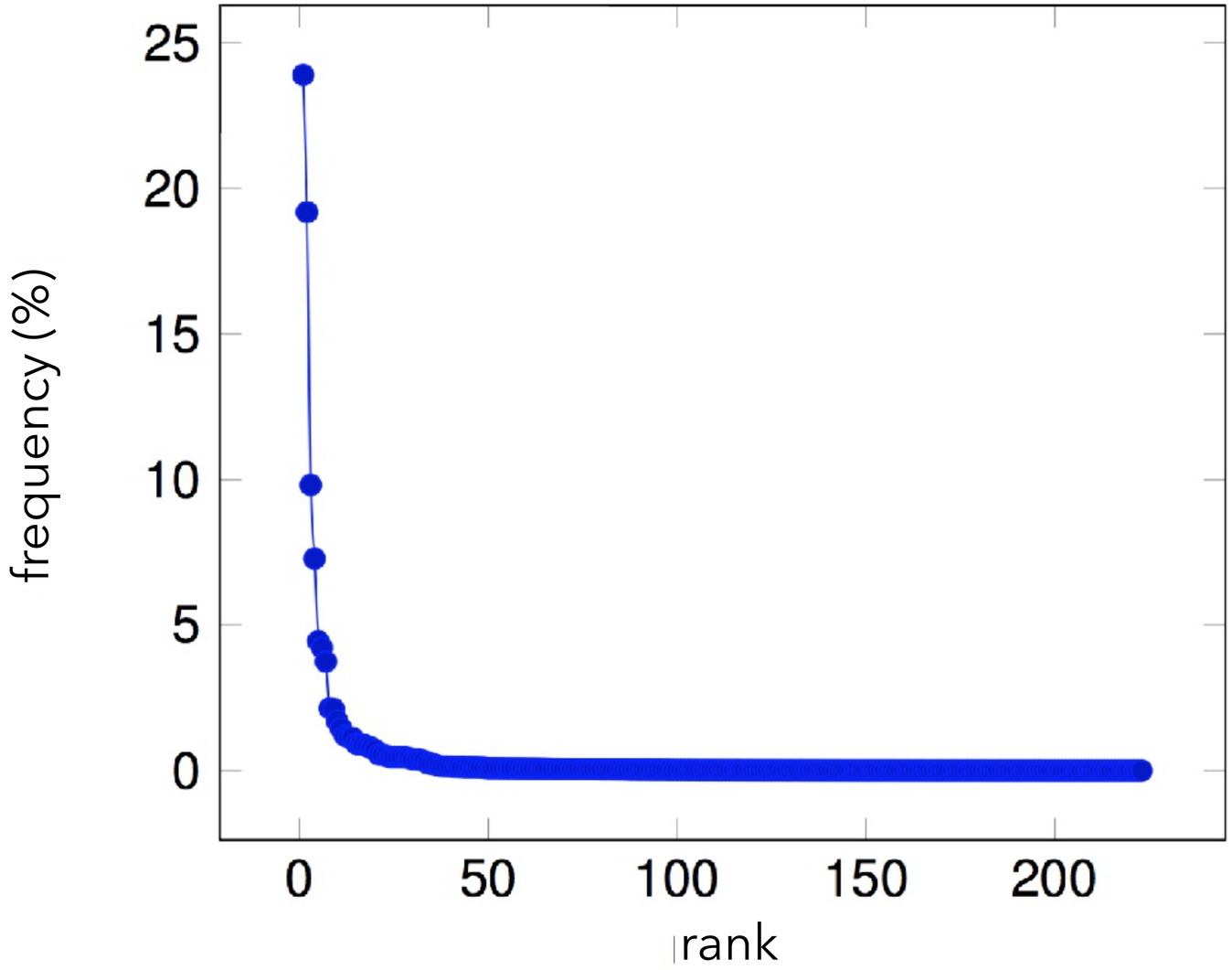


Source: Wikipedia; data: dumps from October 2015

Zipf's law

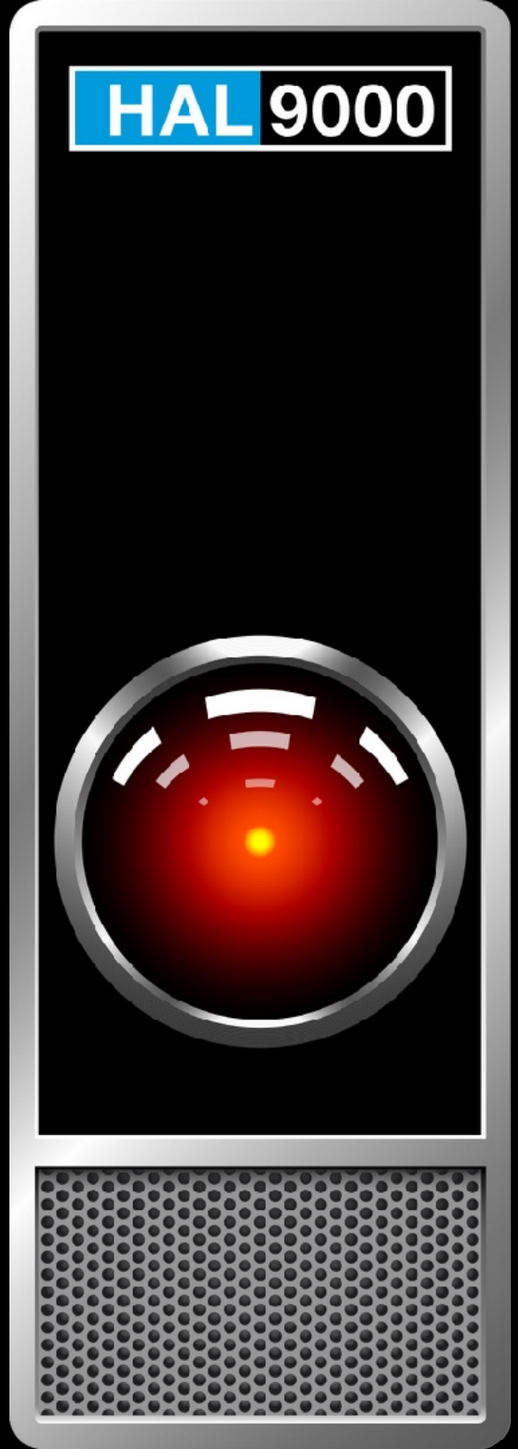
The Zipfian distribution is ubiquitous

Example: frequency of syntactic constructions in a 10,000-word automatically parsed corpus



Source: Éric de la Clergerie

A long-held dream



HAL9000, from Kubrick's *2011, A Space Odyssey*

DALL·E 3 (via ChatGPT, GPT-4, 19/11/2023)