# Compréhension Phonétique du Texte Écrit

**Richard Khoury**

Lakehead
UNIVERSITY

# Contexte

- Traitement statistique du langage naturel
- Nouvelles ressources textuelles
  - Microtext
  - Documents historiques

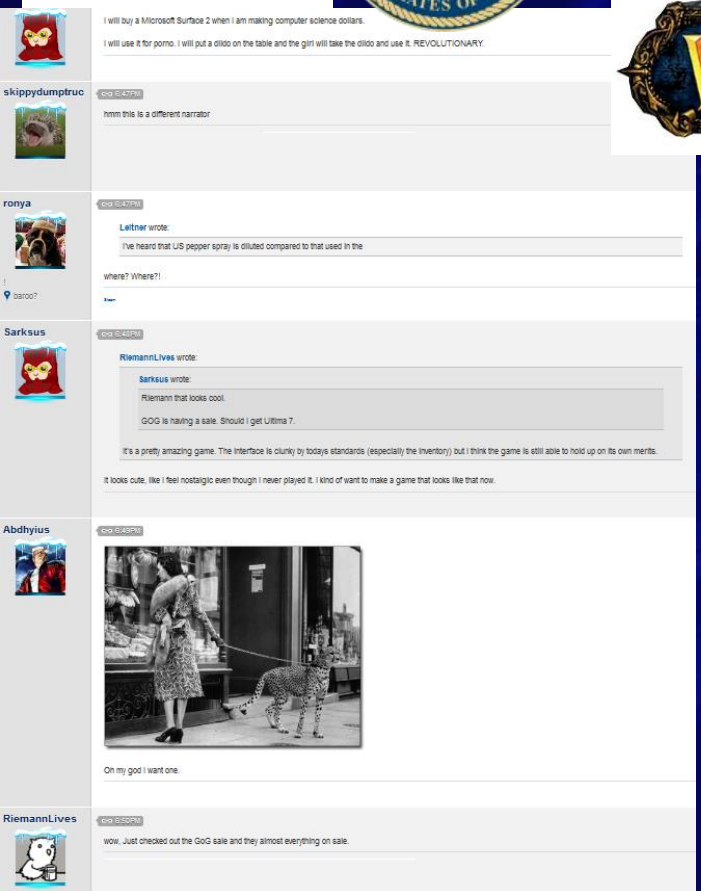
# Microtext

- Un document texte qui est...
  - Très court
    - Quelques mots à quelques phrases
    - Ou même juste un mot
  - Informel
    - Ton de conversation naturelle détendu
    - Plusieurs erreurs d'épellation et de grammaire
  - Semi-structuré
    - Inclus certaines metadata
    - Communément auteur et date
  - [Dela Rosa & Ellen 2009]

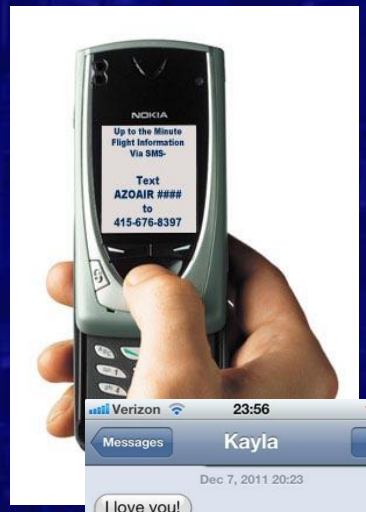
# Pourquoi Étudier le Microtext?

SPAWAR

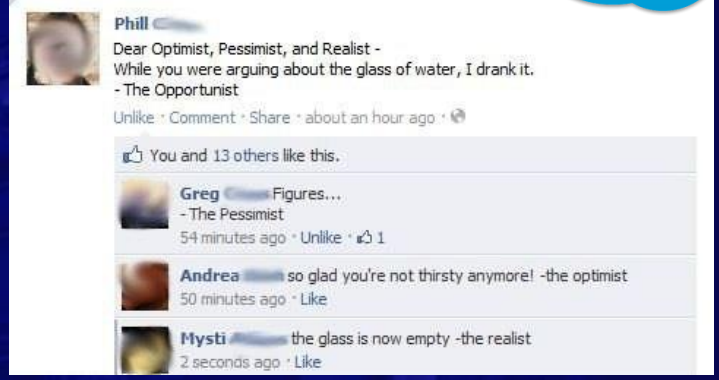














bing

YAHOO!

Baidu 百度

# W31RD SPELIN

- Écriture détendue et informelle
- Aucune vérification pour épellation correcte


- Problème: Outils traditionnels de NLP supposent un langage correct

# Définition du Problème

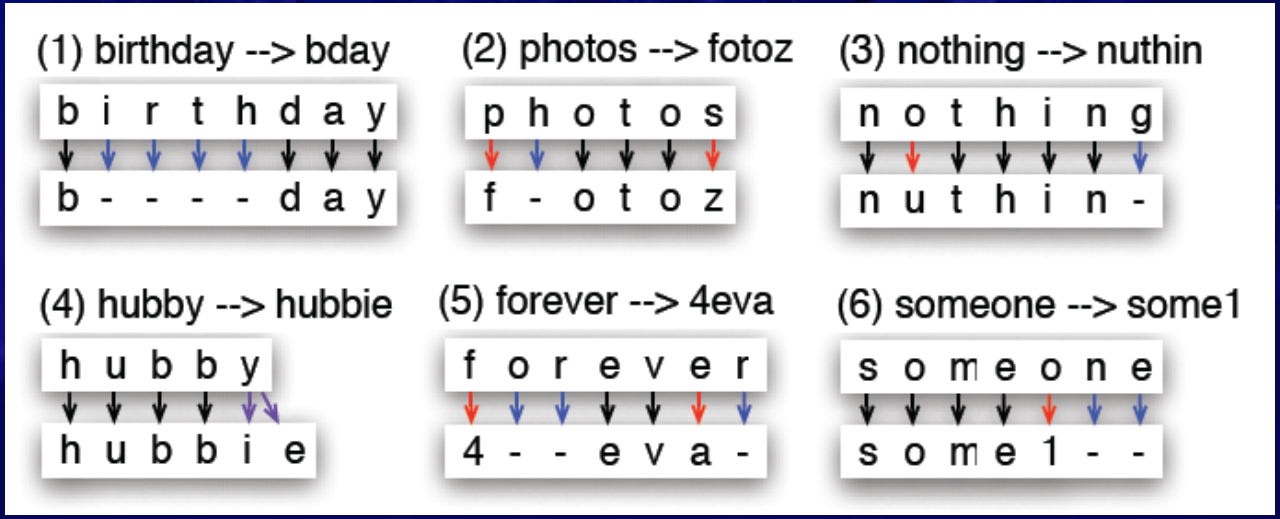
- Normalisation de microtext
  - Déterminer les mots du vocabulaire anglais qui correspondent aux mots hors-vocabulaire du microtext
- Différent de la correction de texte
  - Erreurs de texte dues plus fréquemment à des confusions communes, inversions de lettres, ou à taper la mauvaise touche sur le clavier
- Différent de la compréhension d'acronymes
  - LOL, IMHO, OMG, etc.
  - Ne sont pas des mots mal écrits

# Définition du Problème

- Variations infinies d'épellations
- Innovations constantes
- Cinq types de variations [Liu, et al. 2011]
  - Abréviation
    - together → tgthr
  - Substitution phonétique
    - together → 2gether
  - Substitution graphémique
    - together → t0g3th3r
  - Variation stylistique
    - together → togeda
  - Répétition de lettre
    - together → togetherrr
  - Cas ambigus
    - together → togeter
  - Cas mélangés
    - together → 2gthr

# Approches Alternatives

- Dictionnaire construit de paires contextuelles [Han et al. 2013]
- Algorithme de traduction probabiliste [Pennell & Liu 2014]
- Règles de substitution de lettres [Liu et al. 2011]



# Compréhension Phonétique

- Auteur écrit des mots hors-vocabulaire qui sont phonétiquement similaires aux mots anglais pour être reconnus par le lecteur
- Développer un logiciel qui « lit » les mots phonétiquement puis les apparient aux mots similaires en anglais
- Problème: comment aller du mot écrit à la prononciation?

# Compréhension Phonétique

- Corpus d'entrainement de paires de mots-IPA de Wiktionary

```
{{also|Computer}}
==English==
{{commons}}
{{wikipedia}}
[[Image:Science museum 025 adjusted.jpg|thumb|An electronic computer (circa early 1980s).]]

===Etymology===
From {{suffix|compute|er|lang=en}}.

===Pronunciation===
* {{a|UK}} {{IPA|/kəmˈpju:tə/|lang=en}}
* {{audio|En-uk-computer.ogg|Audio (UK)|lang=en}}
* {{a|US}} {{IPA|/kəmˈpjutɚ/|lang=en}}
* {{audio|en-us-computer.ogg|Audio (US)|lang=en}}
* {{rhymes|u:tə(r)|lang=en}}

===Noun===
{{en-noun}}

# {{context|now|_|rare|chiefly|_|historical|lang=en}} A [[person]] [[employ]]ed to perform [[computation]]s; one who [[compute]]s. {{defdate|from 17th c.}}
#* '''1927''', J. B. S. Haldane, ''Possible Worlds and Other Essays'', p. 173
#*: Only a few years ago Mr. Powers, an American '''computer''', disproved a hypothesis about prime numbers which had held the field for more than 250 years.
#* '''2003''', {{w|Bill Bryson}}, ''A Short History of Nearly Everything'', BCA, p. 116:
#*: One Harvard '''computer''', Annie Jump Cannon, used her repetitive acquaintance with the stars to devise
```

a multilingual free encyclopedia

**Wiktionary**

['wɪkʃənrɪ] *n.*, a wiki-based Open Content dictionary

Wikt ['wɪ kənɪ]

Main Page
Community portal
Preferences
Requested entries
Recent changes
Random entry
Help
Donations
Contact us

Tools
What links here
Related changes
Upload file
Special pages
Printable version
Permanent link
Page information
Cite this page
Add definition

In other projects
Wikipedia

Visibility
Show derived terms
Show related terms
Show translations
Show other boxes
Show conjugation

Entry | Discussion | Citations

## computer

*See also:* **Computer**

### English [edit]

**Etymology** [edit]

From *compute* + *-er*.

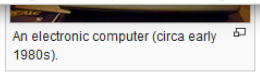
**Pronunciation** [edit]

- (*UK*) IPA(key): /kəmˈpju:tə/
- Audio (UK) ▶ 0:00 MENU
- (*US*) IPA(key): /kəmˈpjutɚ/
- Audio (US) ▶ 0:00 MENU
- Rhymes: -u:tə(r)

**Noun** [edit]

**computer** (*plural* **computers**)

1. (*now rare*, *chiefly historical*) A person employed to perform computations; one who computes. [from 17th c.] [quotations ▼]
2. by restriction, a male computer, where the female computer is called a computress
3. A programmable electronic device that performs mathematical calculations and logical operations, especially one that can process, store and retrieve large amounts of data very quickly; now especially, a small one for personal or home use employed for manipulating text or graphics, accessing the Internet, or playing games or media. [from 20th c.]



An electronic computer (circa early 1980s).

**Quotations** [edit]

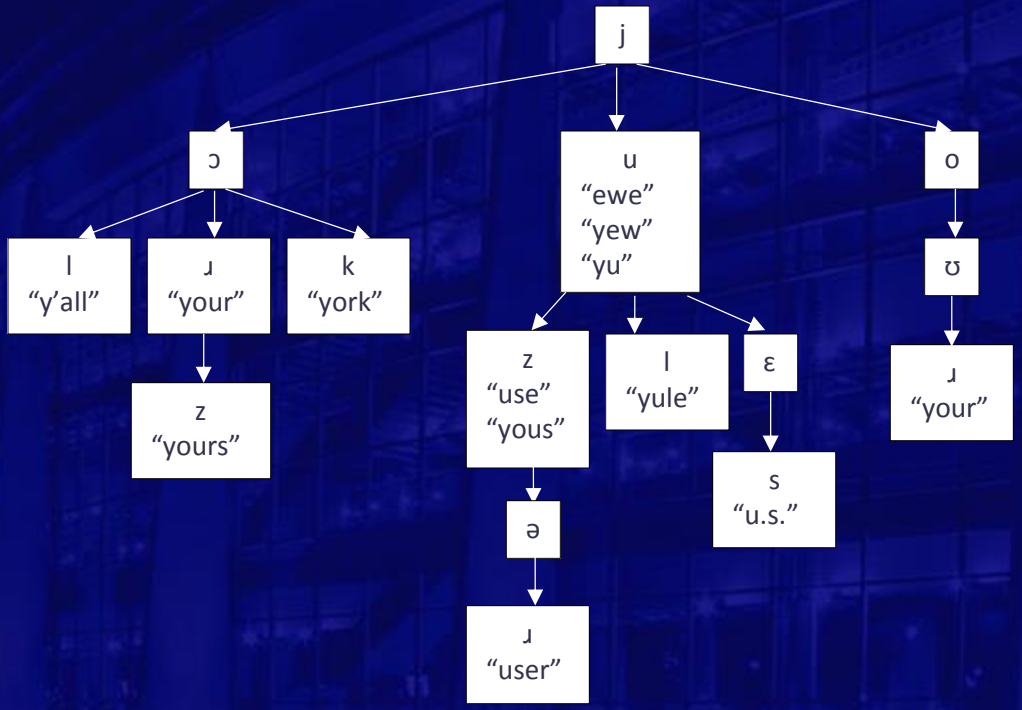
- For usage examples of this term, see the citations page.

# Compréhension Phonétique

- Corpus d'entrainement: 37,500 prononciations de 30,368 mots
- 151 symboles IPA
- Correspondance de (ensemble de) lettre(s) à (ensemble de) symbole(s)
- Calcul des probabilités de correspondances lettres-à-symboles (probabilité globale)

# Compréhension Phonétique

- Construire un arbre radix de l'anglais

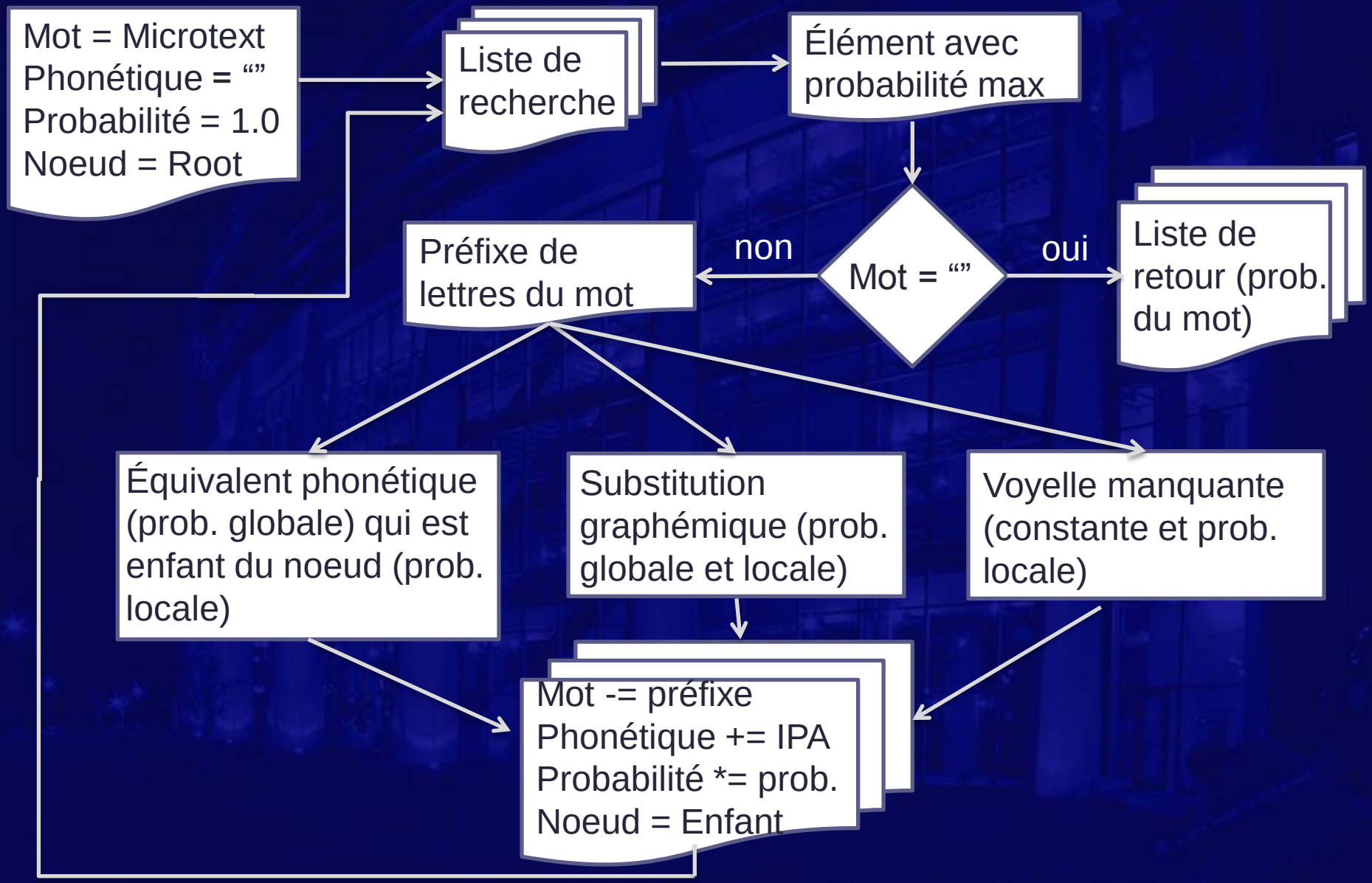



- Calculer probabilité d'un enfant dans chaque chemin (probabilité locale)

# Trouver les Mots Similaires

- Recherche par coût uniforme (suivant)
- Liste de sons similaires
  - Wiktionary a des prononciations très spécifiques
  - “about” = /əˈbaʊt/, /əˈbæʊt/, /əˈbʌʊt/, /əˈbɛʊt/
  - Grouper les sons similaires ensembles
- Probabilité d’utilisation de mots en anglais
- Ensemble de substitutions graphémiques
  - 0-O, 1-I,L, 3-E, 4-A, 7-T

# Trouver les Mots Similaires



# Résultats

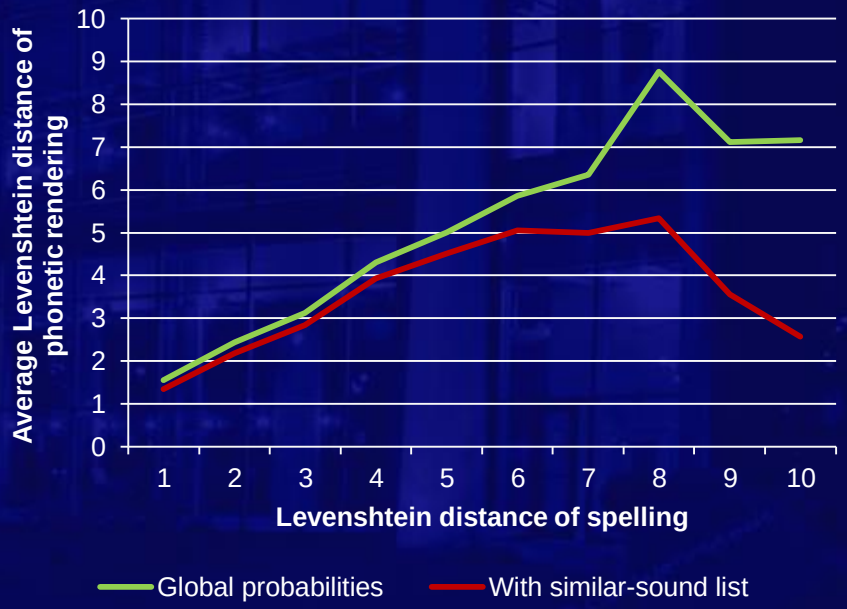
- Corpus de test: 2608 paires de mots de Twitter [Liu, et al. 2011]
- Composition:

| | |
|---|---|
| Abréviation | 806 |
| Substitution phonétique | 130 |
| Substitution graphémique | 58 |
| Variation stylistique | 820 |
| Répétition de lettres | 641 |
| Mélanges de types | 153 |

| | |
|---|---|
| som1 | someone |
| putin | putting |
| havinqq | having |
| 5top | stop |
| ur | your |
| eeeeven | even |
| merecal | miracle |
| wot | what |

# Résultats

- Question 1: Est-ce que l'interprétation phonétique nous aide?
  1. Calculer la distance Levenshtein de l'épellation
  2. Utiliser la probabilité globale pour avoir une interprétation phonétique
  3. Calculer la distance Levenshtein de l'interprétation phonétique
  4. Grouper par distance d'épellation et moyenner
- Réponse: oui
  - Augmentation linéaire initiale malgré l'utilisation de 151 symboles au lieu de 32 caractères
  - Stabilisation aux groupes de distance élevés
  - Augmentation plus limitée lorsqu'on utilise la liste de sons similaires

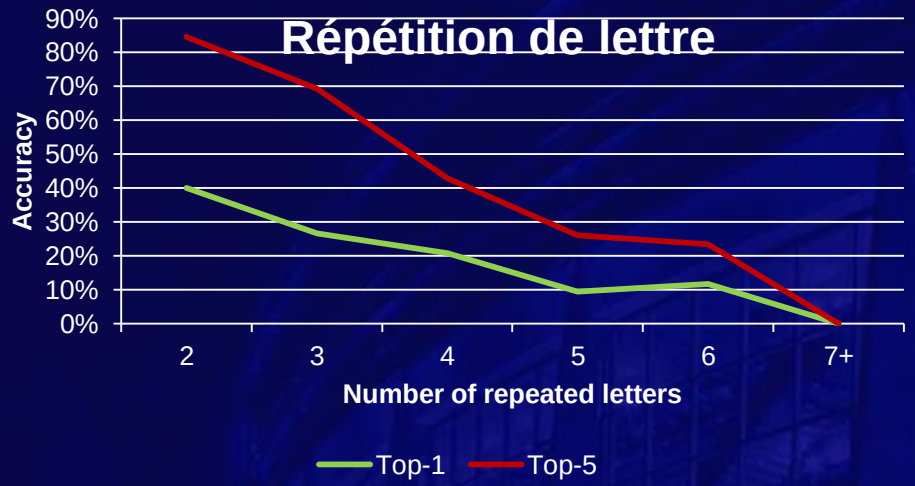


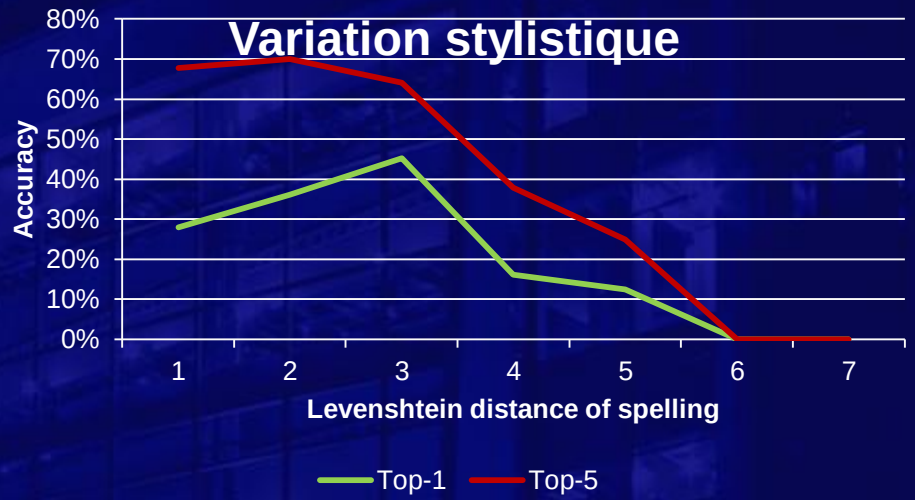
# Résultats

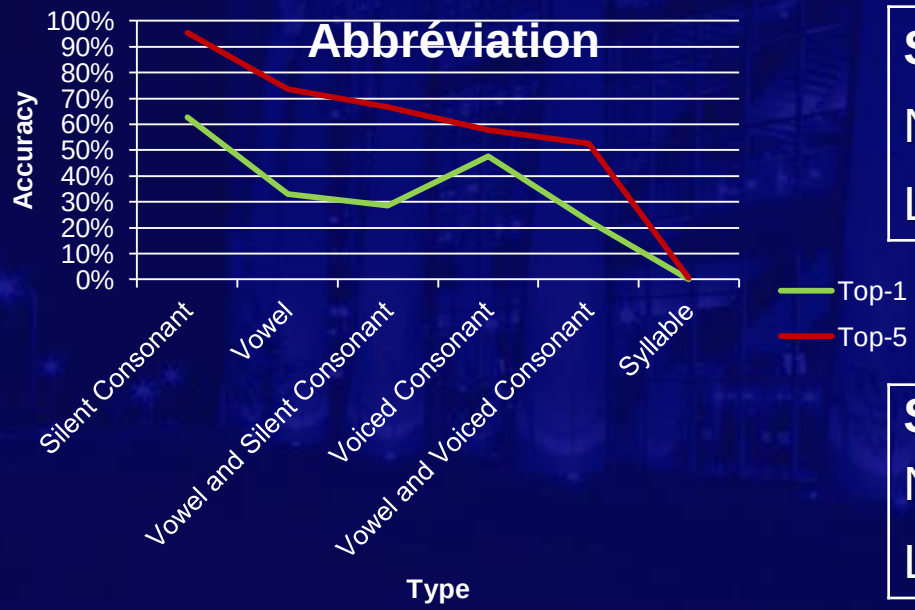
- Question 2: est-ce utile pour la normalisation?
  1. Obtenu top-1 et top-5 mots plus probables pour chaque mot hors-vocabulaire
  2. Étudié types et sous-types de normalisation
- Réponse: oui
  - Écart de 20% à 30% entre top-1 et top-5: les probabilités doivent être ajustées
  - Fonctionne mieux lorsque l'hypothèse de base est respectée

| | Top-1 | Top-5 |
|---|---|---|
| Moyenne globale | 30.2% | 59.7% |
| Abréviation | 29.0% | 52.6% |
| Substitution phonétique | 53.8% | 78.5% |
| Substitution graphémique | 29.3% | 58.6% |
| variation stylistique | 31.6% | 65.7% |
| Répétition de lettres | 28.1% | 62.2% |
| Mélanges de types | 17.9% | 38.4% |

# Résultats




| **Substitution graphémique** | Top-1 | Top-5 |
|---|---|---|
| Nombre à lettre | 45% | 85% |
| Lettre à lettre | 21% | 44% |

| **Substitution phonétique** | Top-1 | Top-5 |
|---|---|---|
| Nombre à son | 18% | 32% |
| Lettre à son | 61% | 88% |

# Moyen Anglais

- Microtext est l'exemple le plus récent d'écrire phonétiquement… mais pas le premier
- Moyen Anglais
  - Angleterre Médiévale du 12th au 15th siècle
  - Avant les dictionnaires et l'orthographe standard
  - Auteurs utilisent plusieurs épellations pour un mot
  - Alphabet inclut lettres runiques: 3 (yogh), Þ (thorn)
  - Substitution du I et Y
  - Substitution du U et V

# Moyen Anglais

16. **Maria.** Þou goddis aungell, meke and mylde,
Howe sulde it be, I the praye,
That I sulde consayve a childe
Of any man by nyght or daye.
I knawe no man þat shulde haue fyled
My maydenhode, the sothe to saye;
With-outen will of werkis wilde,
In chastite I haue ben ay.

17. **Ang.** The Halygast in þe sall lighte,
Hegh vertue sall to þe holde,
The holy birthe of the so bright,
God sonne he sall be calde.
Loo, Elyzabeth, þi cosyne, ne myght
In elde consayue a childe for alde,
Þis is þe sexte moneth full ryght,
To hir þat baran has ben talde.

18. **Maria.** Thou aungell, blissid messanger,
Of goddis will I holde me payde,
I love my lorde with herte dere,
Þe grace þat he has for me layde.
Goddis handmayden, lo! me here,
To his wille all redy grayd,
Be done to me of all manere,
Thurgh thy worde als þou hast saide.

# Moyen Anglais

- Interprétation phonétique devrait permettre de reconnaitre différentes épellations du même mot
- Mais pas de Wiktionary médiéval

# Compréhension Phonétique

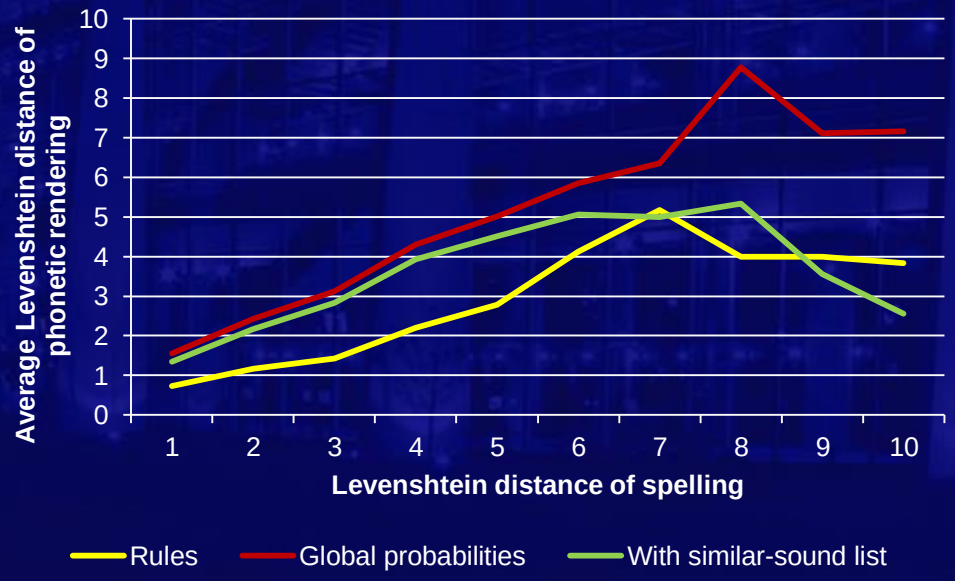
- Liste de 31 symboles de sons en anglais
- 110 règles de prononciation de lettres selon le contexte
- Transforme les mots en séquences phonétiques

si 'a' suivi par 'i' ou 'y': son ← A
sinon si 'a' précédé par 'o': son ← O
sinon si 'a' précédé par 'e' et suivi par 'u': son ← O
sinon si 'a' précédé par 'e': son ← E
sinon si  'a' suivi par consonne suivi par 'e': son ← A
sinon si 'a' suivi par 'u': son ← O
sinon si 'a' suivi par 'r' or 'l' final, optionnellement suivi par 's': son ← 0
sinon: son ← a

| | |
|---|---|
| Silent letter | 0 |
| Short A | a |
| Long A | A |
| Regular B | b |
| Soft C, Soft S | s |
| Hard C | k |
| Regular SH | C |
| Regular D | d |
| Short E | e |
| Long E | E |
| Regular F | f |
| Hard G | g |
| Soft G, Regular J | j |
| Audible H | h |
| Short I | i |
| Long I | I |
| Regular L | l |
| Regular M | m |
| Regular N | n |
| Short O | o |
| Long O | O |
| Regular P | p |
| Regular R | r |
| Regular T | t |
| Short U | u |
| Long U | U |
| Regular V | v |
| Regular W | w |
| Regular OW | W |
| Regular YE | y |
| Regular Z, Hard S | z |

# Compréhension Phonétique

- Question 1: Est-ce que ça marche?
- Refait le test avec 2608 paires de mots de Twitter
  - Trouve la similarité des mots mieux que les probabilités de Wiktionary



# Compréhension Phonétique

- Question 2: Est-ce que ça peut reconnaitre les formes du même mot sans confondre des mots similaires?
- Corpus de tests d'homophones (aye/eye, core/corps, theme/team) et de prochephones (bother/brother, doze/daze, mass/maze)
- Précision:
  - 53% sur les homophones
  - 90% sur les prochephones

# Définition du Problème

- Question 3: Pourquoi manipuler le Moyen Anglais?
- Les documents médiévaux sont étudiés depuis des siècles
  - Mais sans ordinateurs
  - “Étudiés” signifie “lus et résumés manuellement”
  - Analyse logicielle peut révolutionner le domaine

# Définition du Problème

- Cycle de York
  - Pièces de théâtre médiévales sur l'histoire biblique du monde
  - 47 pièces, 381 personnages, 13,307 lignes de dialogue
- Certains personnages utilisent de l'allitération
- Pourquoi?
  - Différents auteurs? [Reese 1951]
  - Signaler des discours importants et émotionnels? [Epp 1989]
  - Démarquer les personnages méchants? [Johnston 1993]

[Scene I, *Pilate's Hall.*]

1. Pil. VNdir þe ryallest roye of rente and renowne,
Now am I regent of rewle þis region in reste,
Obeye vnto bidding bud busshoppis me bowne,
And bolde men þat in batayll makis brestis to breste. 4
To me be-taught is þe tent þis towre begon towne,
For traytoures tyte will I taynte, þe trewþe for to triste,
The dubbyng of my dingnite may noȝt be done downe,
Nowdir with duke nor dugeperes, my dedis are so dreste. 8
My desire muste dayly be done
With þame þat are grettest of game,
And þer agayne fynde I but fone,
Wherfore I schall bettir þer bone. 12
But he þat me greues for a grume,
Be-ware, for wystus I am.

# Définition du Problème

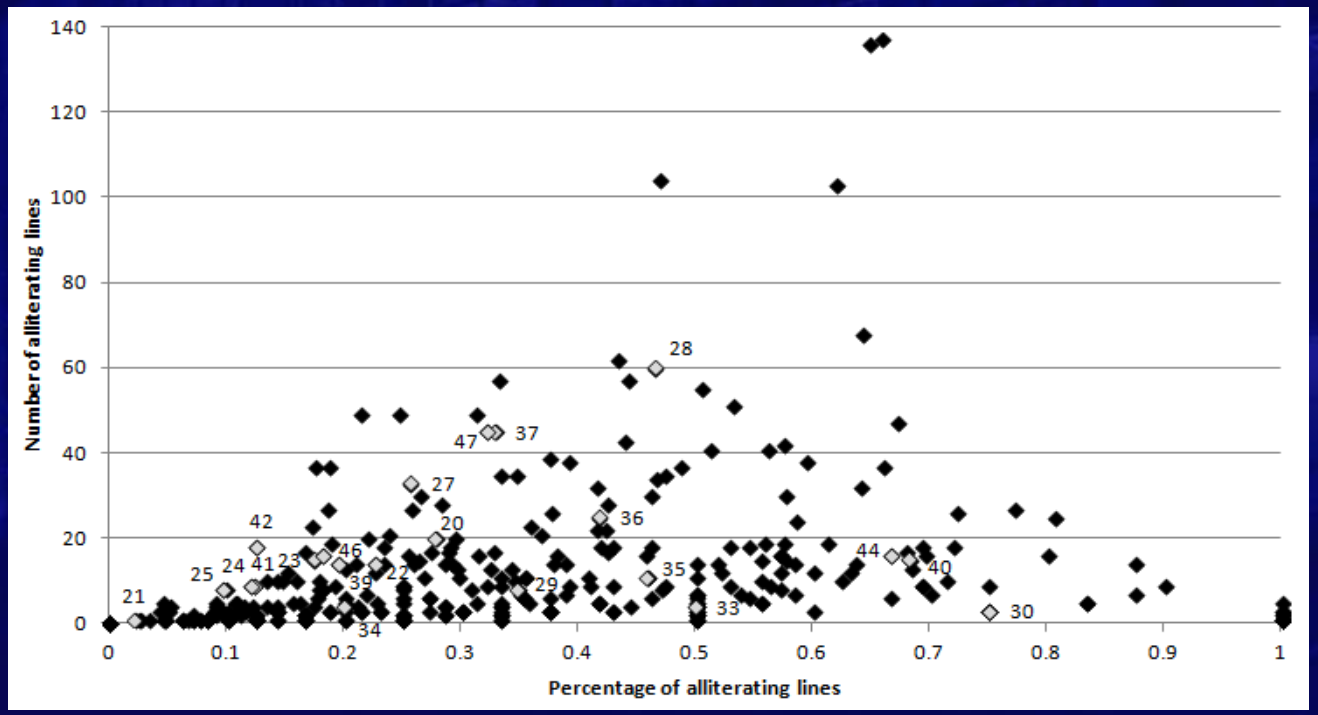
- Avec une interprétation phonétique des mots, il est facile de trouver l'allitération
- Conditions ("mot" exclut mots vides)
  - (son plus commun unique de la ligne) & (utilisé dans au moins trois mots de la ligne) & (utilisé dans au moins la moitié des mots de la ligne)
  - (son plus commun unique de la ligne) & (utilisé dans exactement deux mots de la ligne) & (utilisé dans tous les mots de la ligne)

# Résultats

- Quantification
  - Nombre de lignes avec allitération d'un personnage
  - Proportion de lignes avec allitération d'un personnage
- Créer un nuage de points de l'utilisation d'allitération des personnages

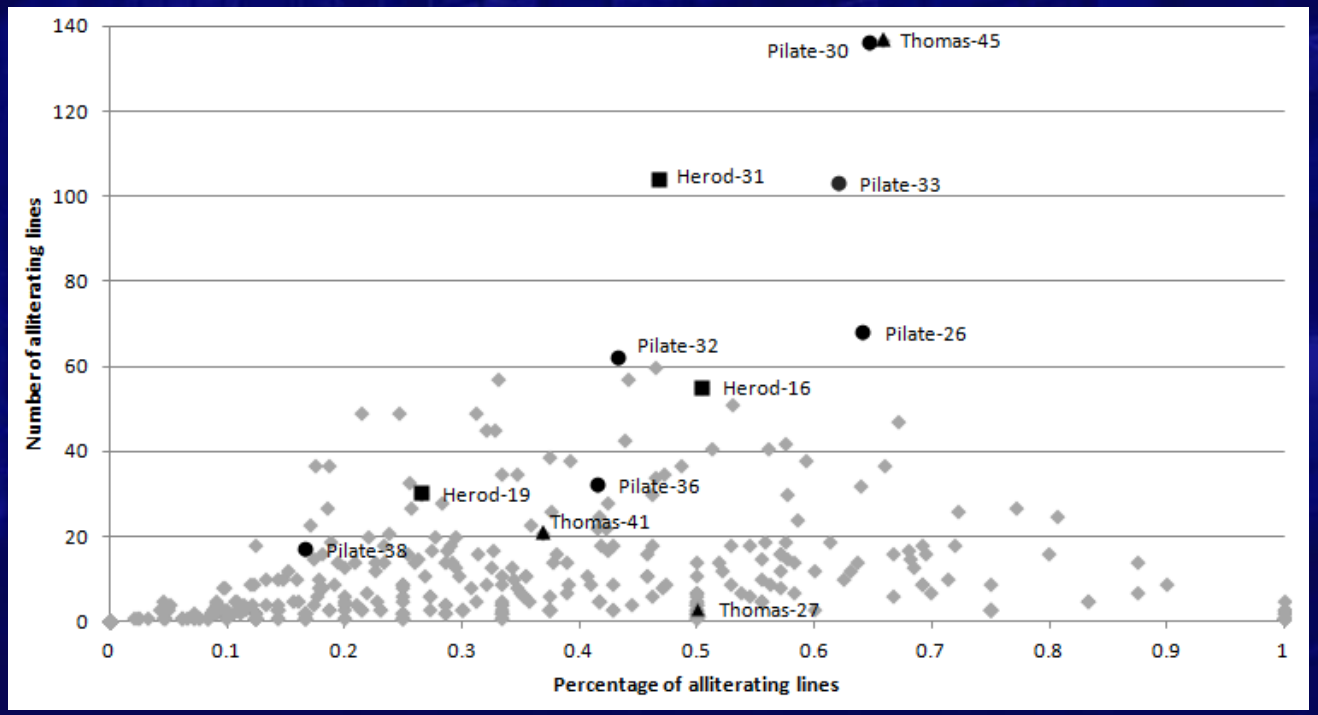
# Résultats

- Allitération omniprésente dans les pièces
  - 28 sur 381 personnages n'utilisent pas d'allitération
  - Moyenne: 11 lignes / 34% des lignes
- Même individu utilise différentes quantités d'allitération dans différentes pièces



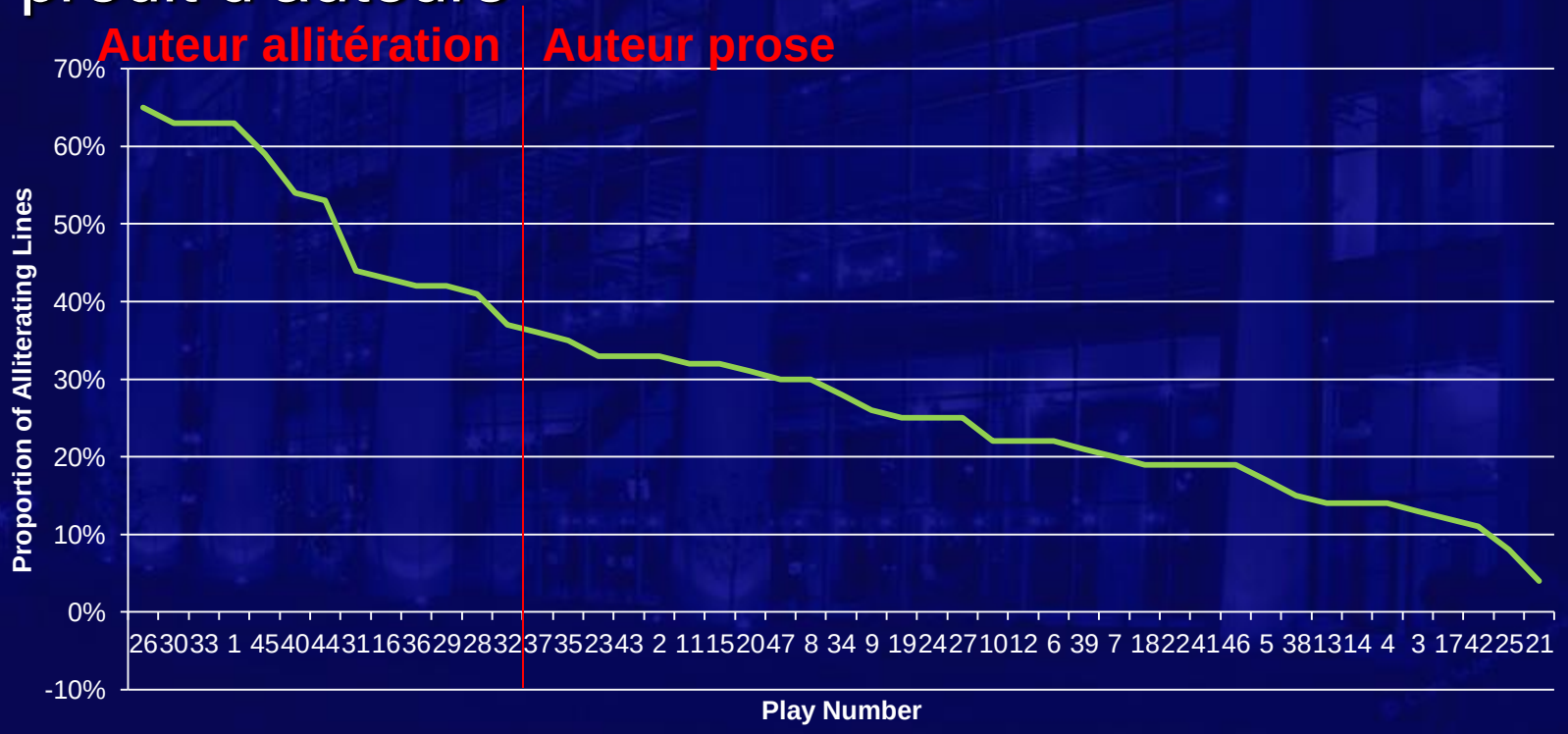
# Résultats

- Théorie 1: allitération marque personnages méchants [Johnston 1993]
- Observation: Les bons et les méchants (et les ambigus) utilisent tous de l'allitération



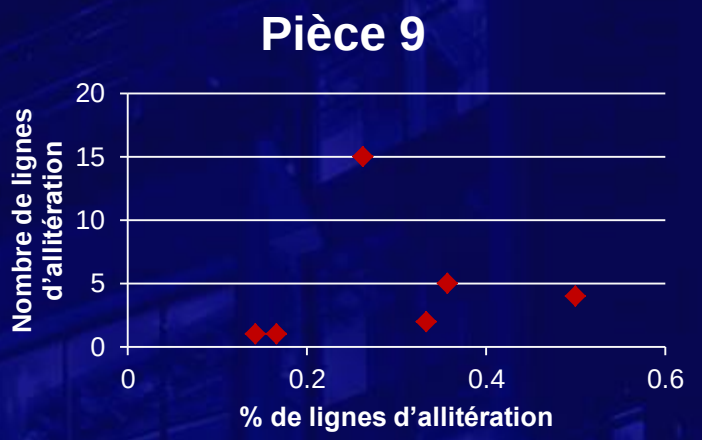
# Résultats

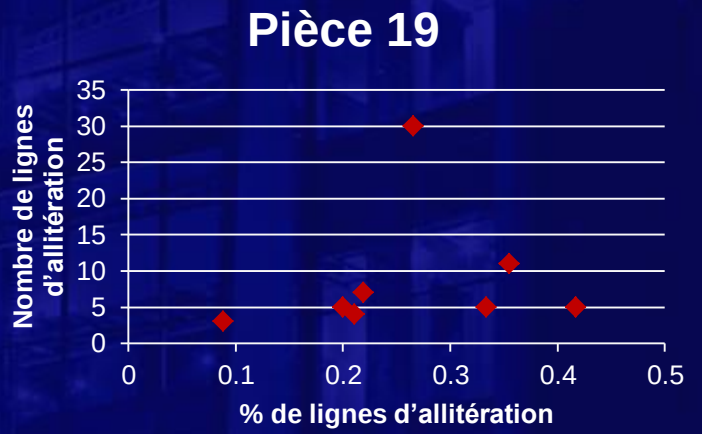
- Théorie 2: allitération indique deux auteurs différents [Reese 1951]
- Observation: Changement d'utilisation d'allitération ne correspond pas au changement prédit d'auteurs

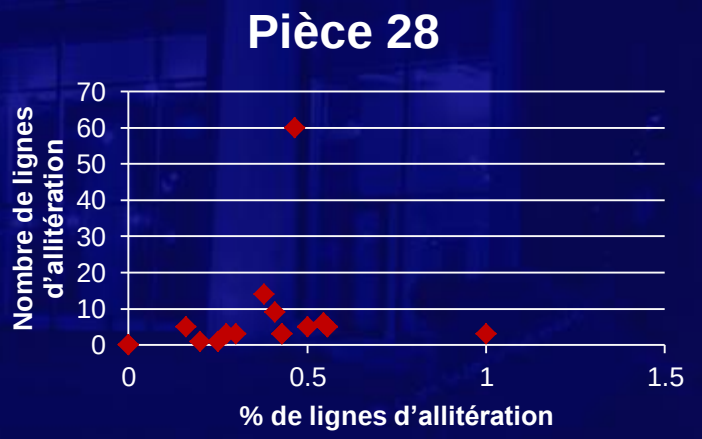


# Résultats

- Allitération par pièce offre une nouvelle perspective
  - 30 pièces ont un personnage avec utilisation d'allitération aberrante
  - 11 pièces ont un duo de personnages aberrants
  - 1 pièce a un trio
  - 4 pièces n'ont pas de personnages aberrants



# Résultats

- Le théâtre médiéval était une activité auditive
  - Les gens allaient “écouter une pièce” [Beadle 2000]
- 47 pièces étaient présentées en une journée sur 47 scènes avec 47 ensembles d’acteurs
  - Identifier les personnages importants dans chaque pièce était un défi pour les spectateurs [Johnston 1993]
- Nouvelle théorie: allitération indique 1-3 personnages importants dans chaque pièce

# Conclusions

- Les gens semblent avoir l'instinct d'écrire comme ils parlent
- Un logiciel qui peut comprendre le texte phonétiquement pour "entendre" ce qui est écrit peut être très utile

# Conclusions

- Deux méthodes pour comprendre phonétiquement le texte écrit
  - Probabilités apprises d'exemples de Wiktionary
  - Règles de prononciation
- Deux domaines d'application
  - Moderne: Normalisation de microtext
  - Historique: Analyse de documents médiévaux

# Références


- Beadle, R., “Verbal Texture and Wordplay in the York Cycle”, *Early Theatre*, 3, 2000, pp. 167-84.
- Dela Rosa, K. and Ellen, J., “Text classification methodologies applied to micro-text in military chat”, *Proceedings of the International Conference on Machine Learning and Applications*, 2009, pp. 710-714.
- Epp, G. P. J., “Passion, pomp and parody: Alliteration in the York plays”, *Medieval English Theatre*, 11:1, 1989, pp. 150-161.
- Han, B., Cook, P., Baldwin, T., “Lexical Normalization for Social Media Text”, *ACM Transactions on Intelligent Systems and Technology*, 4:1, 2013, article 5.
- Johnston, A. F., “The Word made flesh: Augustinian elements in the York cycle”, *The Centre and its Compass: Studies in Medieval Literature in Honor of Professor John Lyerle,* Kalamazoo, MI: Western Michigan University Press, 1993, pp. 225-46.
- Liu, F., Weng, F., Wang, B., and Liu, Y., “Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision”, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, 2011, pp. 71-76.
- Pennell, D. L., and Liu, Y., “Normalization of Informal Text”, *Computer Speech & Language*, 28:1, 2014, pp. 256–277.
- Reese, J. B. “Alliterative verse in the York cycle”, *Studies in Philology*, 48:3, 1951, pp. 639-668.

# Merci!