La linguistique pour le TAL : au service de la simplexité ?

Gaël Lejeune
Université de Caen (GREYC) & Université de Nantes (LINA)
https://lejeuneg.users.greyc.fr/

Strasbourg, 13 novembre 2015

Plan

Simplicité et complexité

Comment les faire cohabiter ?

La **complexité** d'un système provient souvent :

- De l'offuscation, fortuite ou non (pyrotechnie formelle) ;
- De la spécialisation et de la diversification.

La **simplicité** ce peut-être :

- Un manque conceptuel (simplisme) ;
- Un ajustement des moyens et des fins (rasoir d'Ockham).

> *« Simplexity therefore stands for a **balance** between the growing complexity of daily life and our own personal satisfaction. In order to attain this state, we have to **stop always striving to make optimal decisions** »*
>
> Peter Wipperman

Le rapport avec le TAL ?

La citation qui fâche

« Chaque fois que j'enlève un linguiste de mon équipe, mon système se porte mieux. »

Frederick Jelinek (Traitement Automatique du Langage Parlé)

Comment je l'interprète

- Problème communicationnel
- Définition des étapes (moyens) : modèle
- Adaptation aux objectifs (fins) : évaluation

- Modèle "par étapes" VS modèle interprétatif
- Intérêt pour le contenu ou pour l'effet transmis ?

La tradition réductionniste en TAL

Une recette élégante et pratique

Le découpage en unités minimales est commode, on y retrouve :

- les domaines de la linguistique (morphologie, syntaxe) ;
- des concepts des Langages (grammaire, décomposition) ;
- du paradigme de la programmation dynamique ;
- et plus généralement l'idée de **chaîne de traitement** ;
- observables recomposés à partir des unités minimales ;
- à l'opposé d'une approche **holiste**.

Tout problème complexe est-il soluble dans le **réductionnisme ?**, la décomposition vaut-elle simplification ?

La tradition réductionniste en TAL

Une recette élégante et pratique

Le découpage en unités minimales est commode, on y retrouve :

- les domaines de la linguistique (morphologie, syntaxe) ;
- des concepts des Langages (grammaire, décomposition) ;
- du paradigme de la programmation dynamique ;
- et plus généralement l'idée de **chaîne de traitement** ;
- observables recomposés à partir des unités minimales ;
- à l'opposé d'une approche **holiste**.

Tout problème complexe est-il soluble dans le **réductionnisme ?**, la décomposition vaut-elle simplification ?

La tradition réductionniste en TAL

Une recette élégante et pratique

Le découpage en unités minimales est commode, on y retrouve :

- les domaines de la linguistique (morphologie, syntaxe) ;
- des concepts des Langages (grammaire, décomposition) ;
- du paradigme de la programmation dynamique ;
- et plus généralement l'idée de **chaîne de traitement** ;
- observables recomposés à partir des unités minimales ;
- à l'opposé d'une approche **holiste**.

Tout problème complexe est-il soluble dans le **réductionnisme ?**, la décomposition vaut-elle simplification ?

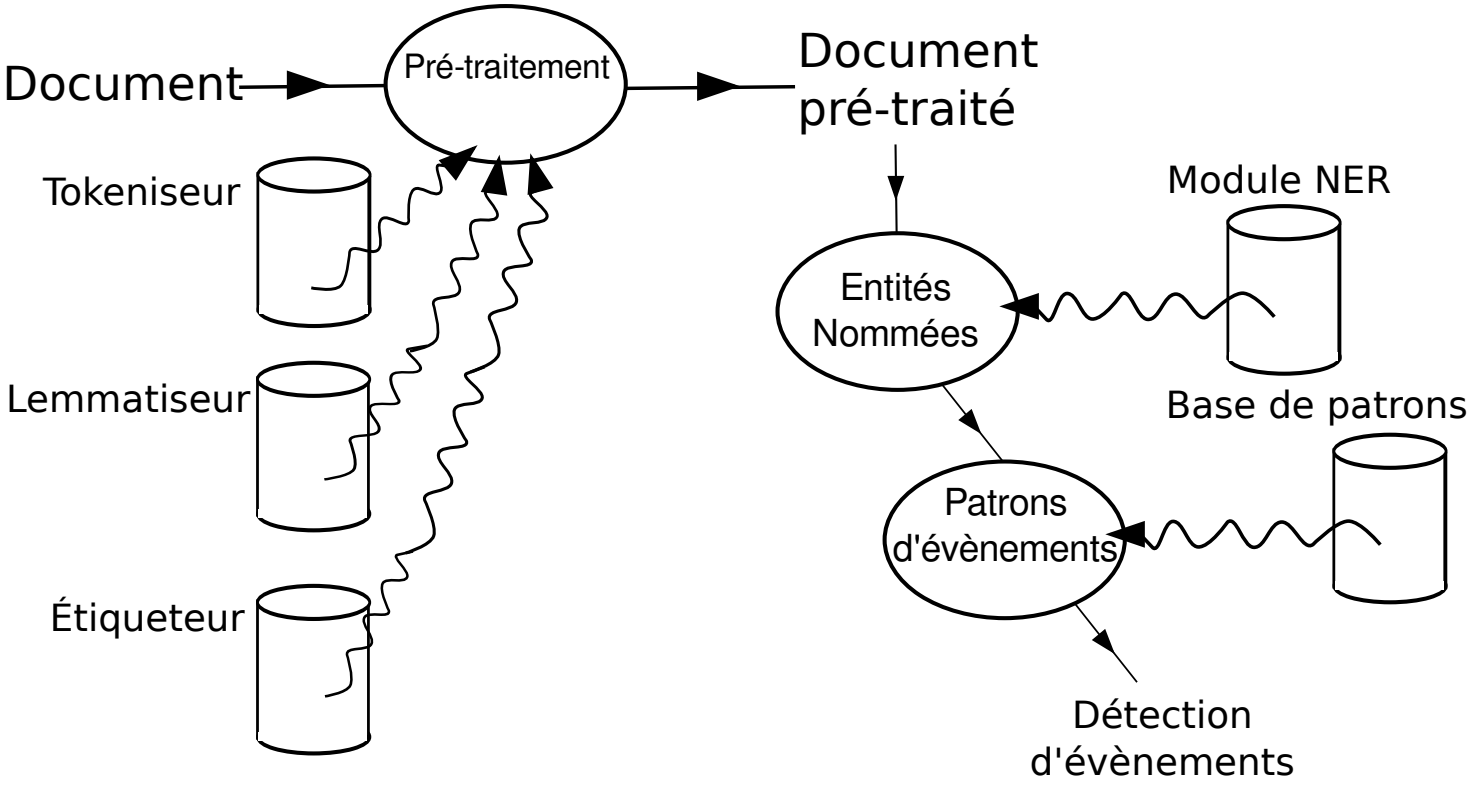
Veille Épidémiologique Multilingue

Exploitation de la presse en ligne

Détecter des évènements le plus rapidement possible

- **Analyse de la langue locale** (signaux « faibles »)
- La réactivité est liée à la **couverture**

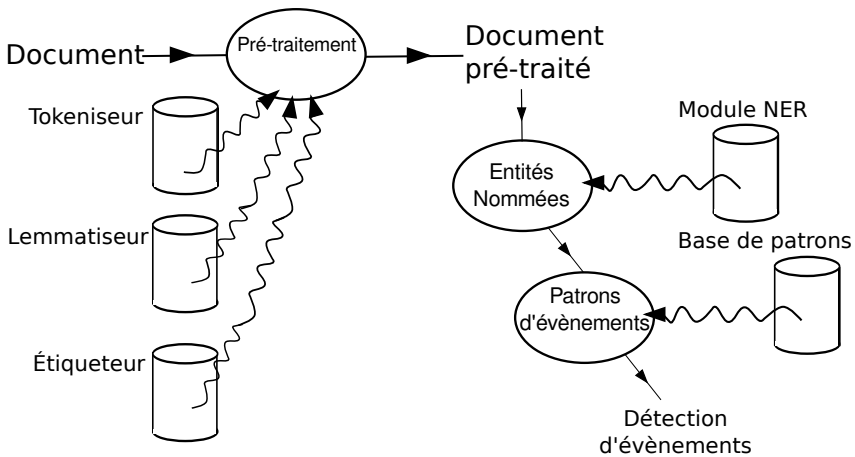
Exemple d'approche réductionniste

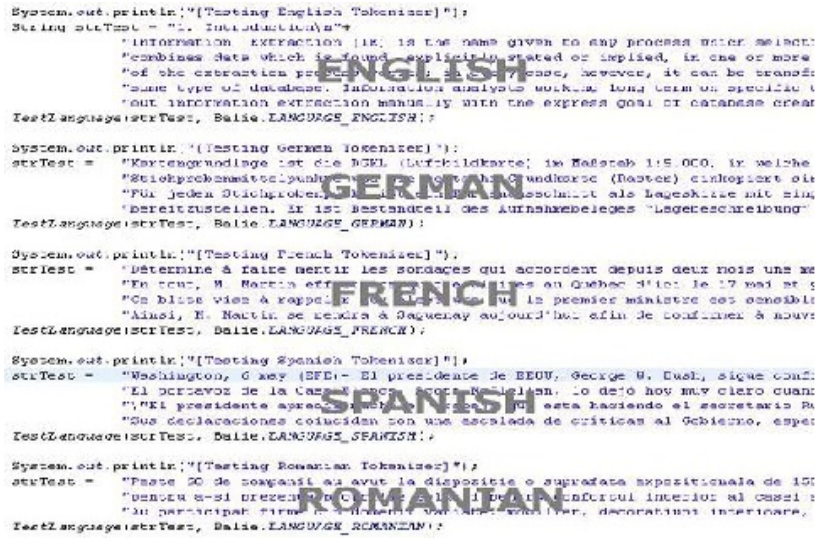


Pré-traitements dépendants de la langue

Analyse coûteuse

Une approche réductionniste « multilingue »

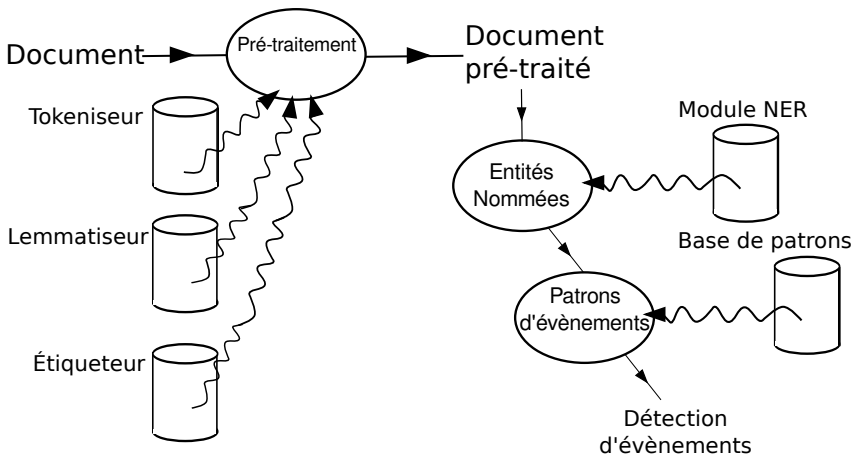


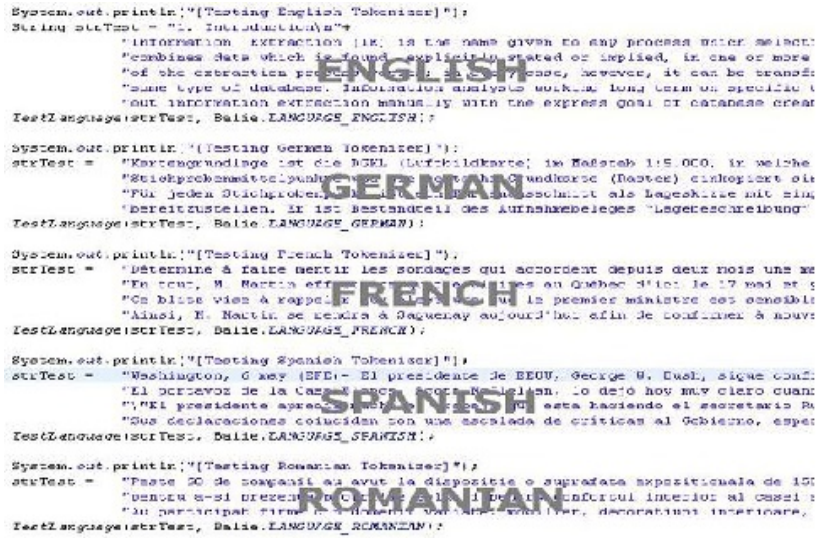


n langues $\Longrightarrow$ *n pipelines*

Approche multi-monolingue$\rightarrow$ à quoi sert le linguiste ?

Une approche réductionniste « multilingue »





n langues $\Longrightarrow$ *n pipelines*

Approche multi-monolingue$\rightarrow$ à quoi sert le linguiste ?

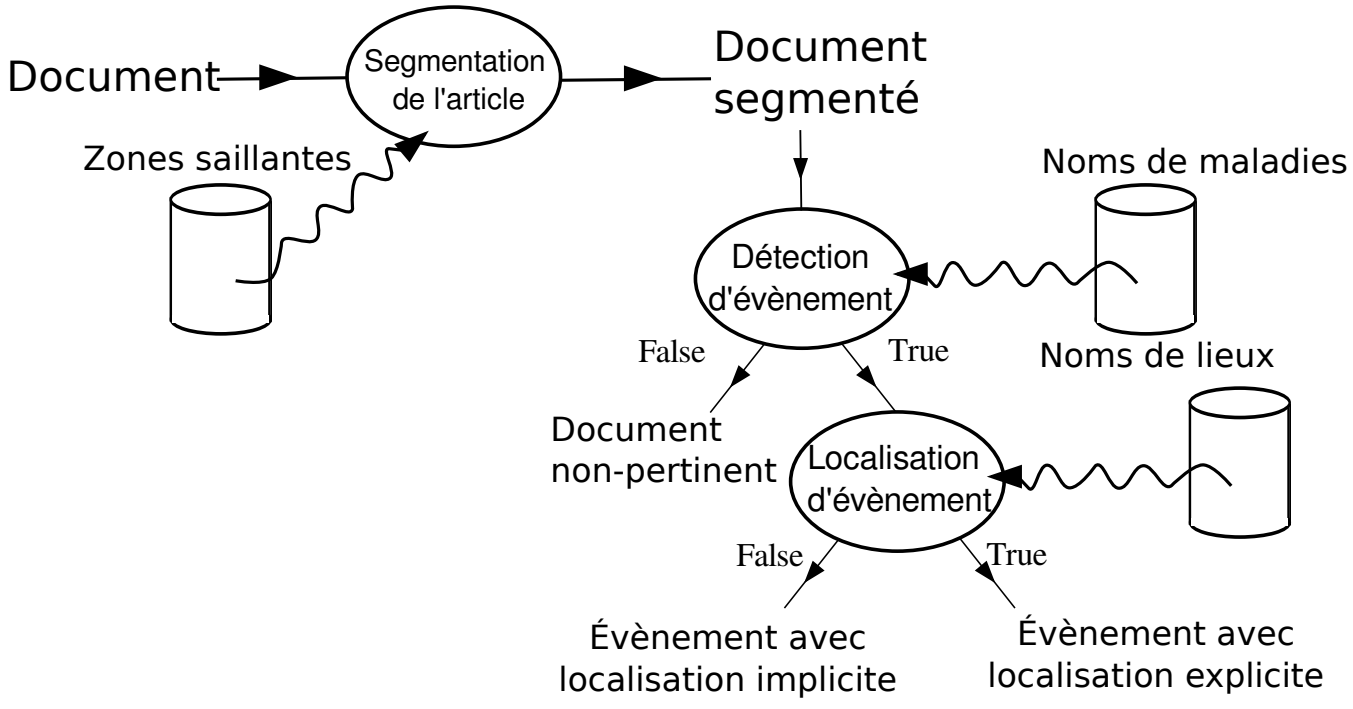
Á quoi sert le linguiste ? Intermède historique

Pourquoi penser simple ?

Livrer **rapidement** un système efficace d'extraction d'évènements dans des textes en français et en espagnol.

- Contrainte de temps
- . . .et de compétences !

L'approche multilingue DANIEL : le *handshake*

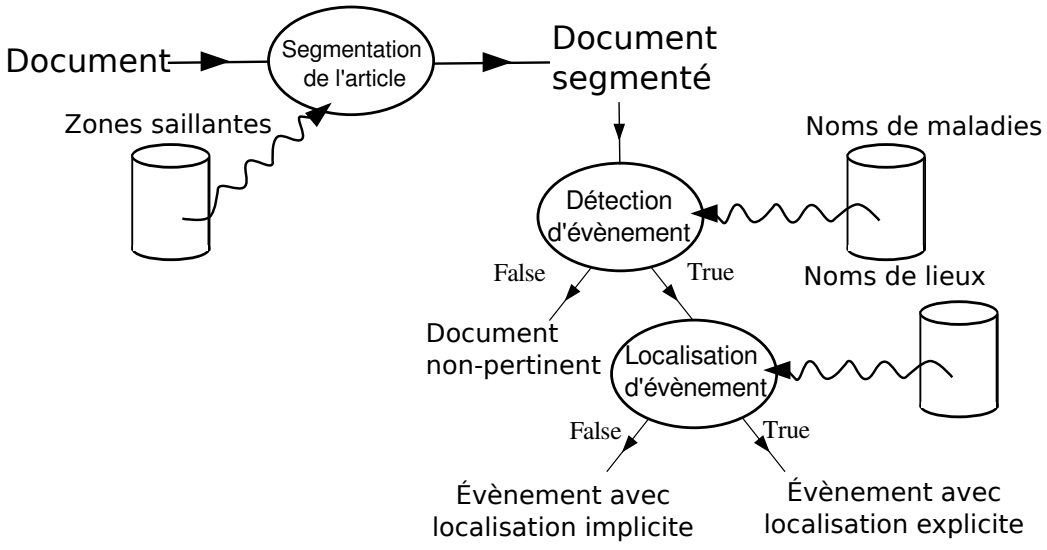


Pré-traitement alingue

Analyse parcimonieuse

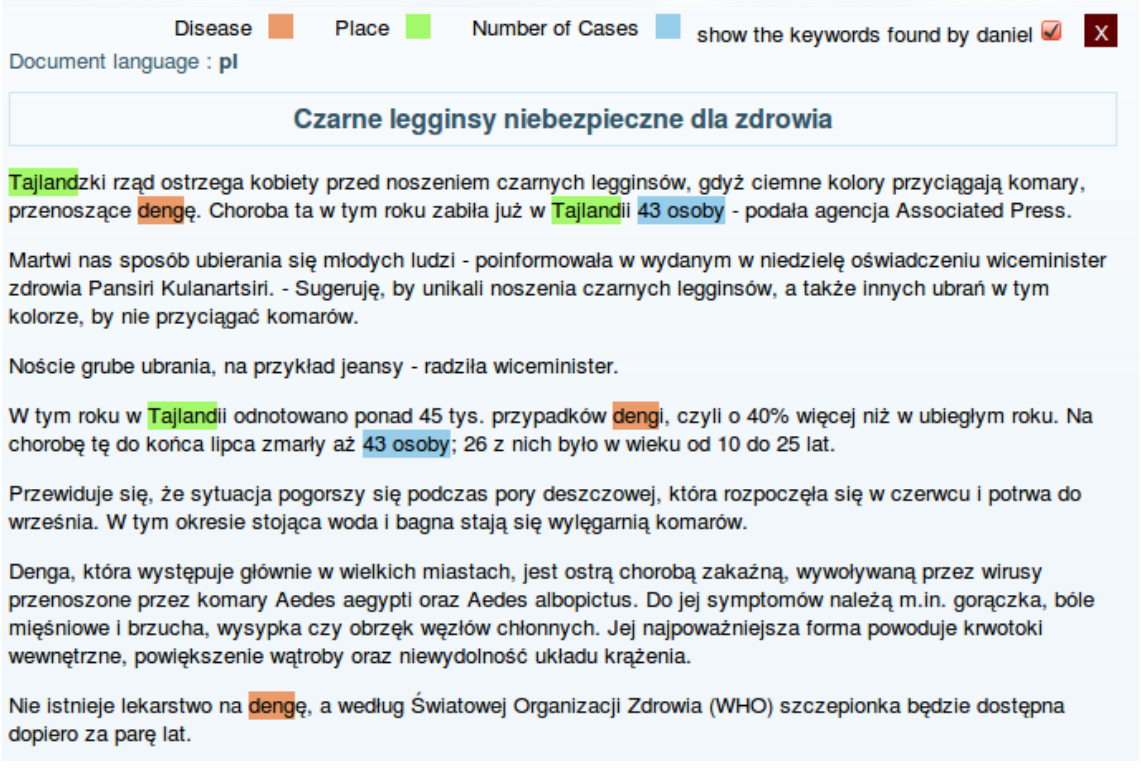
L'approche multilingue DANIEL

Data ANalysis for Information Extraction in any Language



- Propriétés textuelles
- Grain caractère
- Factorisation (réutilisation)
- Parcimonie (coût marginal)

53 langues dont le polonais



Disease Place Number of Cases show the keywords found by daniel X

Document language : **pl**

Czarne legginsy niebezpieczne dla zdrowia

Tajlandzki rząd ostrzega kobiety przed noszeniem czarnych legginsów, gdyż ciemne kolory przyciągają komary, przenoszące dengę. Choroba ta w tym roku zabiła już w Tajlandii 43 osoby - podała agencja Associated Press.

Martwi nas sposób ubierania się młodych ludzi - poinformowała w wydanym w niedzielę oświadczeniu wiceminister zdrowia Pansiri Kulanartsiri. - Sugeruję, by unikali noszenia czarnych legginsów, a także innych ubrań w tym kolorze, by nie przyciągać komarów.

Noście grube ubrania, na przykład jeansy - radziła wiceminister.

W tym roku w Tajlandii odnotowano ponad 45 tys. przypadków dengi, czyli o 40% więcej niż w ubiegłym roku. Na chorobę tę do końca lipca zmarły aż 43 osoby; 26 z nich było w wieku od 10 do 25 lat.

Przewiduje się, że sytuacja pogorszy się podczas pory deszczowej, która rozpoczęła się w czerwcu i potrwa do września. W tym okresie stojąca woda i bagna stają się wylęgarnią komarów.

Denga, która występuje głównie w wielkich miastach, jest ostrą chorobą zakaźną, wywoływaną przez wirusy przenoszone przez komary Aedes aegypti oraz Aedes albopictus. Do jej symptomów należą m.in. gorączka, bóle mięśniowe i brzucha, wysypka czy obrzęk węzłów chłonnych. Jej najpoważniejsza forma powoduje krwotoki wewnętrzne, powiększenie wątroby oraz niewydolność układu krążenia.

Nie istnieje lekarstwo na dengę, a według Światowej Organizacji Zdrowia (WHO) szczepionka będzie dostępna dopiero za parę lat.

Les différentes formes de **Tajland-** et **deng**. . .

. . .sont mises en relation sans analyse grammaticale.

53 langues dont le finnois

Disease Place Number of Cases show the keywords found by daniel X

Document language : **fi**

Kolera tappanut jo yli 3000 Zimbabwessa

Kolera tappanut jo yli 3000 Zimbabwessa

julkaistu 28.01.2009 klo 12:37, päivitetty 28.01.2009 klo 14:10

Eteläafrikkalaisessa Zimbabwessa koleraan kuolleiden määrä on noussut jo yli kolmen tuhannen. Maailman terveysjärjestön WHO:n mukaan kuolleita on 3 028.

Lisäksi yli 57 000 ihmistä on sairastunut. Tiistain jälkeen on rekisteröity 57 uutta kuolemantapausta ja yli 1 500 uutta tartuntaa.

Elokuussa puhjennut epidemia on Afrikan pahin 14 vuoteen.

Zimbabwen presidentti Robert Mugabe on väittänyt, ettei Zimbabwessa enää ole koleraa. Hänen mukaansa viranomaiset ovat saaneet taudin kuriin, ja muut väitteet ovat ulkovaltojen yritystä heikentää hänen asemaansa.

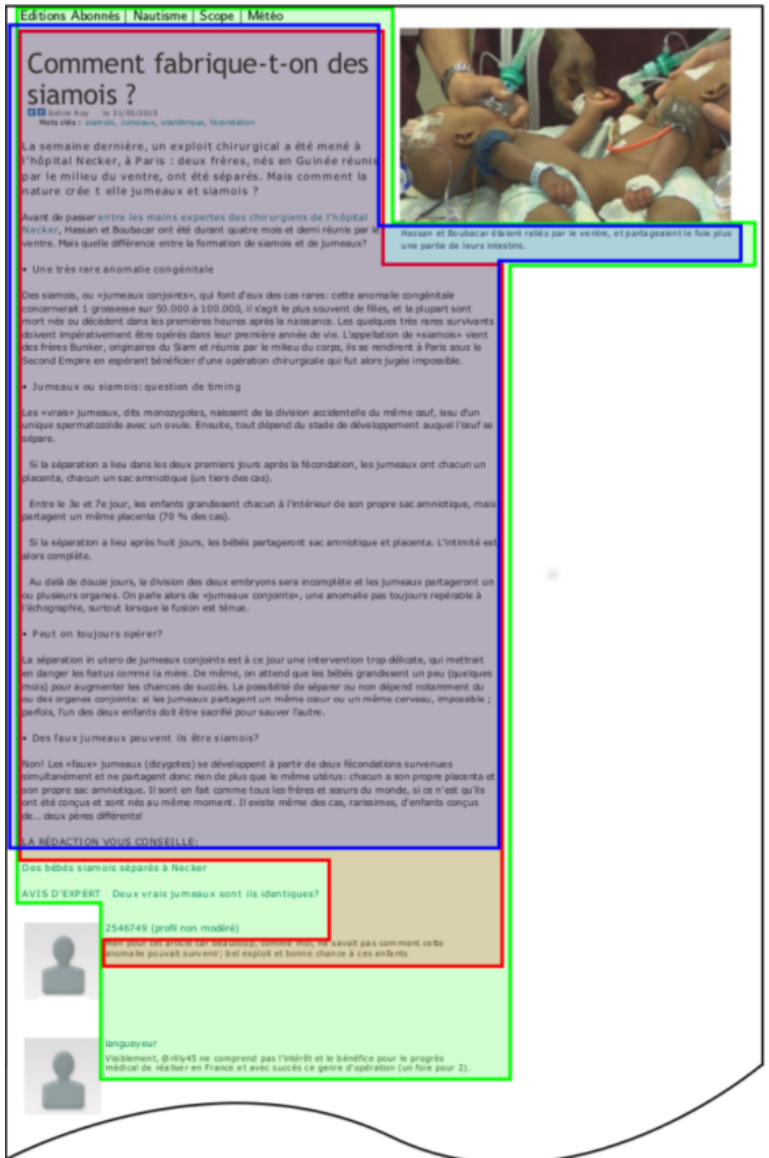
Tauti on levinnyt myös Zimbabwen naapurimaihin, Etelä-Afrikkaan, Botswanaan, Sambiaan ja Mosambikiin.

kolera tarttuu bakteerin saastuttamasta ruoasta tai juomavedestä. Tauti aiheuttaa ripulia, oksentelua ja niistä johtuvan nestehukan. Pahimmassa tapauksessa kolera johtaa kuolemaan.

Reuters, AFP, YLE Uutiset

$\rightarrow$ la complexité des fins par la simplicité des moyens.

Le nettoyage de pages Web



Évaluation par la F-mesure

Text Only (TO) ;

Text and Mark-up (TM)

Caractères

page 2	TO	TM	CAR
BoilerPipe	91,24	94,94	33,33
JustText	86,93	34,68	43,75
NCleaner	68,43	23,26	21,05

Intuitivement JT>BP, NC OK

Comment interpréter ?

Évaluation extrinsèque

Interroger les modalités d'évaluation

- Par le **contenu** (intrinsèque) difficilement interprétable ;
  - Distance d'édition : pertinence ?
  - Qu'est-ce qu'un bon résultat ?
  - Variabilité selon les sources et les langues.
- Par la **tâche** (extrinsèque) : plus adaptée
  - Variations encore plus fortes selon les langues ;
  - Valeurs trompeuses (absolues comme relatives) ;
  - Évaluation du maillon non-détachable de celle de la chaîne.

Un outil en principe simple mais d'évaluation complexe

→La finalité du texte devrait être au centre de l'évaluation

Quelles unités ont valeur terminologique ?

Quel est le sens activé ? Est-ce le sens termininologique ?

Ingrédients utilisés classiquement :

- lemme / étiquettes grammaticales
- voisinage dans la phrase

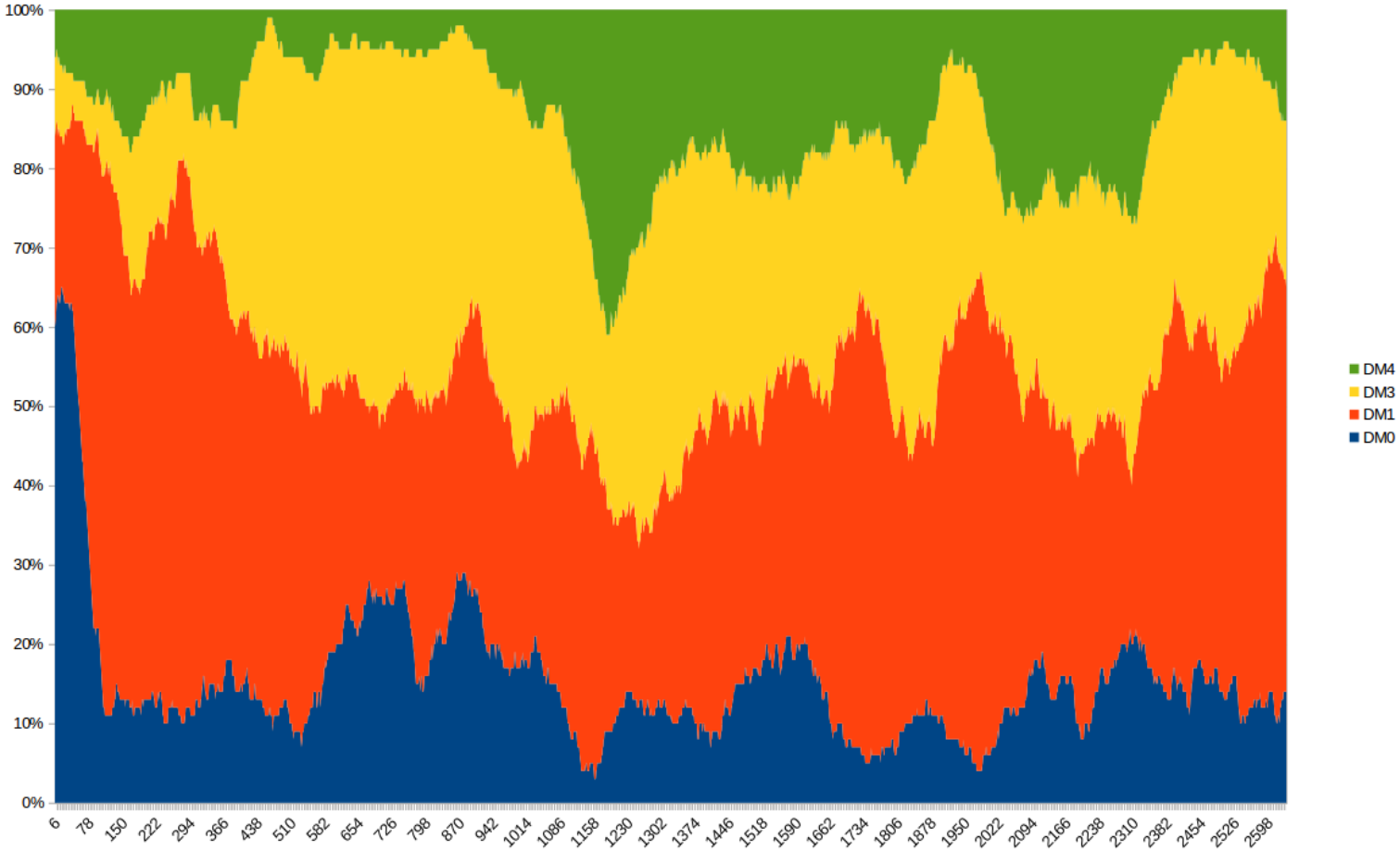
Améliorations possibles de la **recette** :

- trouver la bonne combinaisons d'ingrédients ;
- avoir le bon algorithme de cuisson.

Arrières goûts :

- Interprétabilité (encore) ;
- Généricité ;
- Reproductibilité (selon les langues notamment).

Classification de candidats termes en SHS



Répartition au fil du texte des candidats termes, des moins au plus terminologiques (de bas en haut).

L'écosys–terme (Lejeune & Daille TALN 2015)

Partir des stratégies de lecture du document

Hypothèse : les textes ont une structure grâce à laquelle les termes sont mis en valeur (saillance)

- un algorithme **simple** (2 lignes) :
  - Pour chaque candidat terme ;
  - Calculer sa position par rapport aux balises XML.
- des règles **lisibles** :
  - les termes sont fréquents dans les titres ;
  - mais aussi dans les appels à références. . .
- Plus efficace et moins coûteux (calculs/ressources).

Questionnement sur le réductionnisme en TAL

Comment traiter les données textuelles ?

- Données Non-structurées ou computationnellement opaques (De Busser 2006)
- Peut-on justifier la chaîne de (mauvais) traitement(s) ?
- Peut-on calculer du sens (Seleskovitch 1981, Rastier 1999, Coursil 2000, Beust 2013) ?

Pour aller plus loin sur la question des **observables** :

- Empirisme–rationnalisme (Church Coling 2010)
- « Pauvreté conceptuelle »(F.Yvon Journée ATALA 04/2014)
- **La complexité est masquée plutôt qu'affrontée**

Questionnement sur le réductionnisme en TAL

Comment traiter les données textuelles ?

- Données Non-structurées ou computationnellement opaques (De Busser 2006)
- Peut-on justifier la chaîne de (mauvais) traitement(s) ?
- Peut-on calculer du sens (Seleskovitch 1981, Rastier 1999, Coursil 2000, Beust 2013) ?

Pour aller plus loin sur la question des **observables** :

- Empirisme–rationnalisme (Church Coling 2010)
- « Pauvreté conceptuelle »(F.Yvon Journée ATALA 04/2014)
- **La complexité est masquée plutôt qu'affrontée**

Questionnement sur le réductionnisme en TAL

Comment traiter les données textuelles ?

- Données Non-structurées ou computationnellement opaques (De Busser 2006)
- Peut-on justifier la chaîne de (mauvais) traitement(s) ?
- Peut-on calculer du sens (Seleskovitch 1981, Rastier 1999, Coursil 2000, Beust 2013) ?

Pour aller plus loin sur la question des **observables** :

- Empirisme–rationnalisme (Church Coling 2010)
- « Pauvreté conceptuelle »(F.Yvon Journée ATALA 04/2014)
- **La complexité est masquée plutôt qu'affrontée**

La simplexité dans le TAL en un slide

Interroger le *pipeline* du TAL

- Parcimonie (approche endogène) ;
- Non-compositionnalité, aspects relationnels ;
- Les principes VS les recettes ;
- Interpréter les résultats ou les étapes ;
- Ajuster les moyens et les fins.

La simplexité $\rightarrow$ accepter la simplicité

Merci de votre attention

Aristote *: « Il vaut mieux prendre des principes moins nombreux et de nombre limité »*

Ockham *: « Une pluralité ne doit pas être posée sans nécessité. »*

Thomas d'Aquin *: « ...ce qui peut être accompli par des principes en petit nombre ne se fait pas par des principes plus nombreux. . . »*

Wittgenstein *: « Si un signe n'a pas d'usage, il n'a pas de signification. »*