

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

Gaël Lejeune

Université de Caen (GREYC) & Université de Nantes (LINA)

<https://lejeuneg.users.greyc.fr/>

Strasbourg, 13 novembre 2015



Plan

- 1 La simplicité qu'est-ce que c'est ?
- 2 Simplicité et multilinguisme
- 3 Simplicité et interprétation des résultats
- 4 Simplicité et ambiguïté
- 5 Discussion

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).

*« Simplicity 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 ?

- 1 La simplicité qu'est-ce que c'est ?
- 2 Simplicité et multilinguisme
- 3 Simplicité et interprétation des résultats
- 4 Simplicité et ambiguïté
- 5 Discussion

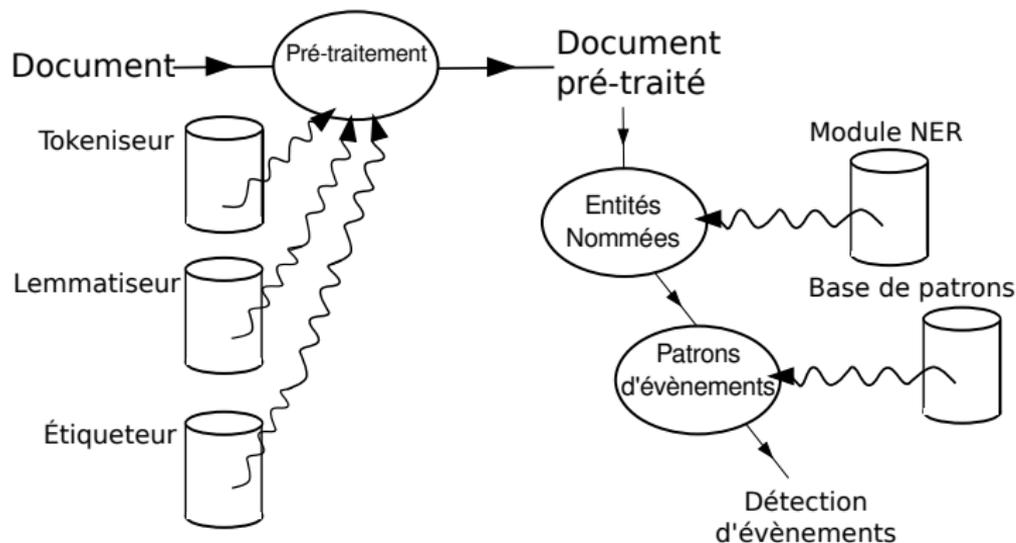
Veille Épidémiologique Multilingue

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**

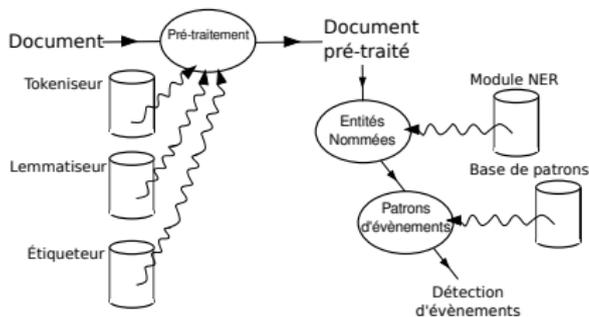
Exemple d'approche réductionniste



Pré-traitements
dépendants de la langue

Analyse coûteuse

Une approche réductionniste « multilingue »



```

System.out.println("[Testing English Tokenizer]");
String s1Text = "1. Introduction";
// INFORMATION EXTRACTION: is it ever given no any process error without
// needless data which is found, applied, stated or implied, in one or more
// for the extraction process? However, it can be assumed
// some type of document. Information extraction module: how can we specify it
// out information extraction module: is with the express goal of database search
TestLanguageTests, Base.LANGUAGE_ENGLISH;

System.out.println("[Testing German Tokenizer]");
// Processing the test text
// "Für jeden Deutschland, der im Jahr 1990 ein Mitglied der Europäischen Union
// wurde, ist die Mitgliedschaft ein Recht der Mitgliedsstaaten."
TestLanguageTests, Base.LANGUAGE_GERMAN;

System.out.println("[Testing French Tokenizer]");
// Détecter et faire sortir les passages qui accordent depuis deux fois un se
// "En fait, le Brest est le plus au Québec d'ici le 17 mai de l'
// "Ce Brest veut à savoir: le Brest est le premier ministre des ministres
// "Ainsi, le Brest de la région de Québec est celui de la région de Québec et non
TestLanguageTests, Base.LANGUAGE_FRENCH;

System.out.println("[Testing Spanish Tokenizer]");
String s1Text = "Washington, D.C. (EEF) - El presidente de EE.UU., George W. Bush, dijo ayer
// "El presidente español, José María Aznar, se dirigió hoy al cónsul de
// "Las declaraciones coinciden con una semana de críticas al Gobierno, espe
TestLanguageTests, Base.LANGUAGE_SPANISH;

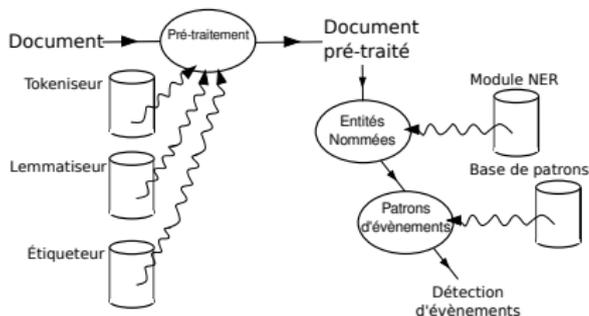
System.out.println("[Testing Romanian Tokenizer]");
String s1Text = "Primele 10 de săptămâni au avut la dispoziție o echipă de specialitate de 100
// "Unul din cele mai importante proiecte de investiții în România este
// "Cu participarea financiară a statului român, dezvoltarea infrastructurii
TestLanguageTests, Base.LANGUAGE_ROMANIAN;

```

n langues $\implies n$ pipelines

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

Une approche réductionniste « multilingue »



```

System.out.println("[Testing English Tokenizer]");
System.out.println("1. Initialisation");
INFO:INFO: Extraction list: in the case given no any previous error method;
Possibles data which is found, applied, stored or updated, in one or more
For the extraction process, however, it can be assumed;
Case type: no elements; Information elements: none; Item: no specific;
OUT: Information extraction with: if with the express goal: of database item
TestLanguageTest: Suite.LANGUAGE_ENGLISH;

System.out.println("[Testing German Tokenizer]");
Series = "Forschungsbildung des IRL (Infrastruktur) im Rahmen des IRL, in welche
Schulungsbildungsbildung im Bereich der Ausbildung (Foster) durchgeführt ist
Für jede Durchführung, die im Bereich der Ausbildung durchgeführt wird, ist
Schulungsbildung, in der Ausbildung, die Ausbildungsbildung (Lernschritte)";
TestLanguageTest: Suite.LANGUAGE_GERMAN;

System.out.println("[Testing French Tokenizer]");
Series = "entendre & faire sentir les braves qui accourent depuis deux fois une se
En cas, le bruit est le bruit de la guerre d'été, le 17 mai de l
Ce bruit est le bruit de la guerre d'été, le 17 mai de l
Ainsi, le bruit de la guerre d'été, le 17 mai de l
TestLanguageTest: Suite.LANGUAGE_FRENCH;

System.out.println("[Testing Spanish Tokenizer]");
Series = "Washington, D key (DIP) - El presidente de DEVO, George W. Bush, sigue con
El proceso de la Unión entre los Estados Unidos, de otro lado, se está con
El presidente de la Unión entre los Estados Unidos, de otro lado, se está con
Que declaraciones coinciden con los resultados de críticas al Gobierno, espe
TestLanguageTest: Suite.LANGUAGE_SPANISH;

System.out.println("[Testing Romanian Tokenizer]");
Series = "Faza de dezvoltare a avut la dispozitie o oportunitate specializata de ID
"Oportunitate de dezvoltare a avut la dispozitie o oportunitate specializata de ID
"Oportunitate de dezvoltare a avut la dispozitie o oportunitate specializata de ID
TestLanguageTest: Suite.LANGUAGE_ROMANIAN;
  
```

n langues \Rightarrow n pipelines

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

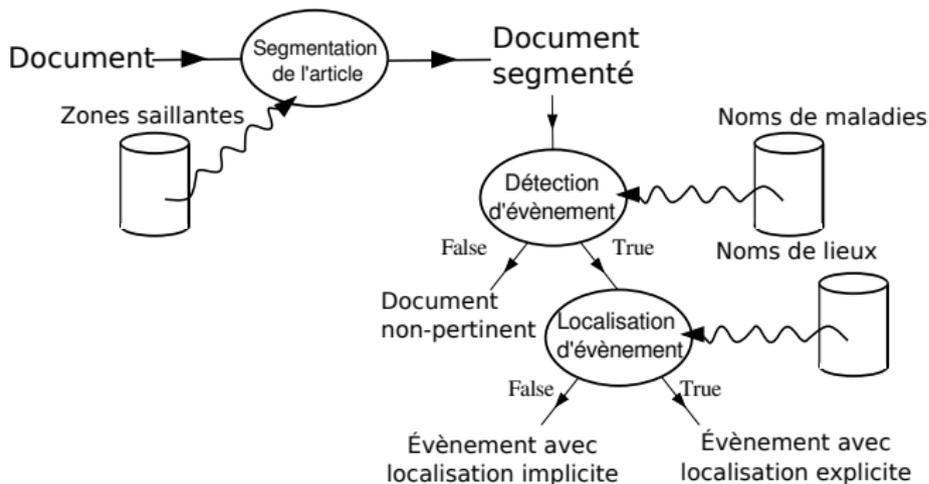
À quoi sert le linguiste ? Interimède historique

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*

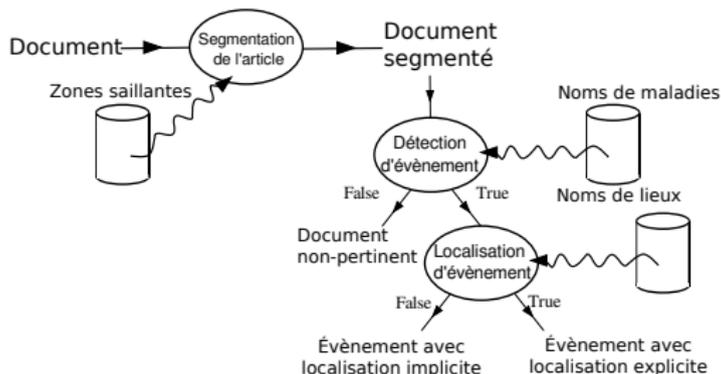


Pré-traitement alingue

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

Document language : pl

Czarne legginsy niebezpieczne dla zdrowia

Tajlandzki rząd ostrzega kobiety przed noszeniem czarnych legginsów, gdyż ciemne kolory przyciągają komary, przynoszące **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 **deng**, 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 potwa 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 przynoszone 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 ✕

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** **kolera**an 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 välttänyt, ettei **Zimbabwe**ssa enää ole **kolera**a. Hänen mukaansa viranomaiset ovat saaneet taudin kurin, ja muut välitteet 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. Pählimmassa tapauksessa **kolera** johtaa kuolemaan.

Reuters, AFP, YLE Uutiset

→ la complexité des fins par la simplicité des moyens.

- 1 La simplicité qu'est-ce que c'est ?
- 2 Simplicité et multilinguisme
- 3 Simplicité et interprétation des résultats
- 4 Simplicité et ambiguïté
- 5 Discussion

Le nettoyage de pages Web

Éditions Abonnés | Navigation | Scope | Menu

Comment fabrique-t-on des siamois ?

100 mots-clés - le 12/06/2010

Paroles, gestes, images, vidéos, sonores

La semaine dernière, un exploit chirurgical a été mené à l'hôpital Necker, à Paris : deux frères, nés en Gascogne du côté du milieu du ventre, ont été séparés. Mais comment la nature a-t-elle créé jumeaux et siamois ?

Avant de passer entre les mains expertes des chirurgiens de l'hôpital Necker, Hassan et Badouar ont été séparés quatre mois et demi dans un ventre. Mais quelle différence entre la formation de siamois et de jumeaux ?

- Une très rare anomalie congénitale

Des siamois, ou « jumeaux congénites », qui font l'objet de cas rares, cette anomalie congénitale survient entre le 20ème et le 300ème jour de la grossesse de l'épouse, et le plus souvent avant même qu'elle ne soit dans les premières heures après la naissance. Les jumeaux nés dans ces circonstances sont généralement deux enfants dans leur propre ventre de mère. L'opération de séparation vient des frères Boukar, originaire du Bénin et décédé par le milieu du corps, et se rendent à Paris pour le second Empire en exploitant l'expertise d'une spécialiste chirurgicale qui fut alors jugée impossible.

- Jumeaux ou siamois: question de timing

Les frères jumeaux, très monozygotes, naissent de la division accidentelle de même œuf, leur développement se fait dans un même placenta, leur développement jusqu'à leur naissance.

Si la séparation a lieu dans les deux premiers jours après la fécondation, les jumeaux ont chacun un placenta, chacun un sac amniotique (qui leur sert de nid).

Entre le 3e et 14e jour, les enfants grandissent chacun à l'intérieur de son propre sac amniotique, mais partagent un même placenta (70 % des cas).

Si la séparation a lieu après huit jours, les bébés partagent un amniotique et placenta. L'opération est alors complexe.

Au-delà de deux jours, la division des deux embryons sera incomplète et les jumeaux partageront un seul placenta unique. On parle alors de « siamois congénites », une anomalie pas toujours compatible à l'ultra-sonographie, surtout lorsque la fusion est totale.

- Peut-on les séparer après ?

La séparation in utero de jumeaux congénites est à ce jour une intervention très difficile, qui consiste à séparer les fœtus comme le main. On réalise, en effet, que les bébés grandissent en peu d'espace dans leur propre sac amniotique. La possibilité de séparer ces deux enfants dépend notamment du lieu des organes congénites et des jumeaux partagent un même cœur ou un même cerveau, intestins, reins, etc. Les deux enfants sont donc destinés à mourir.

- Mais faire jumeaux séparés, de quoi s'agit-il ?

Une fois les frères jumeaux identifiés en développement à partir des deux fécondations séparées (simultanément et les partagent dans leur sac à deux ou à un même œuf), chacun a son propre placenta et son propre sac amniotique. Il s'agit en fait comme tous les frères et sœurs du monde, si ce n'est qu'il y a eu deux ovules et deux spermatozoïdes au lieu d'un seul. Il existe même des cas, rares, où les enfants sont de... deux sexes différents.

LA RÉDACTION VOUS CONSEILLE :

Des vidéos, photos, images et sonores à Necker

VOUS D'ESPÉRER Deux vrais jumeaux sont-ils identiques ?

1/2010 (graphisme studio)

Le contenu de cette page est sous licence Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International. Vous pouvez partager et modifier le contenu de cette page à condition de mentionner le nom de l'auteur et de partager vos modifications.

graphisme

graphisme, développé en partenariat avec l'Institut de la Santé et de la Sécurité pour le Travail (INRS) et l'Institut de la Santé et de la Sécurité au Travail (INRS).



Les frères Boukar, originaire du Bénin et décédé par le milieu du corps, et se rendent à Paris pour le second Empire en exploitant l'expertise d'une spécialiste chirurgicale qui fut alors jugée impossible.

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

- 1 La simplicité qu'est-ce que c'est ?
- 2 Simplicité et multilinguisme
- 3 Simplicité et interprétation des résultats
- 4 Simplicité et ambiguïté
- 5 Discussion

Quelles unités ont valeur terminologique ?

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

Ingrédients utilisés classiquement :

- lemme / étiquettes grammaticales
- voisinage dans la phrase

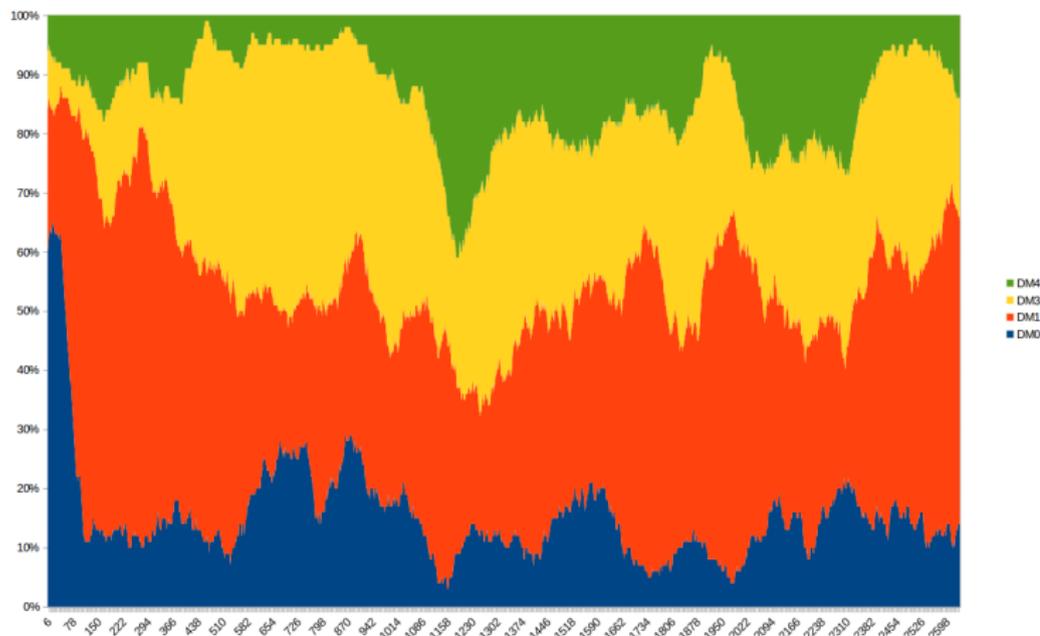
Améliorations possibles de la **recette** :

- trouver la bonne combinaison 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).

- 1 La simplicité qu'est-ce que c'est ?
- 2 Simplicité et multilinguisme
- 3 Simplicité et interprétation des résultats
- 4 Simplicité et ambiguïté
- 5 Discussion

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 simplicité 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 simplicité → 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. »