# Comment mesurer la similarité morphologique dérivationnelle ?

**Nabil Hathout**

CLLE/ERSS
CNRS & Université de Toulouse 2 Le Mirail

Séminaire de l'axe TAL de CLLE/ERSS
14 octobre 2013

# Vue d'ensemble

# Introduction

- Les relations morphologiques s'établissent entre des mots qui partagent en même temps certaines de leurs propriétés sémantiques et phonologiques
- Les propriétés partagées varient d'un couple à l'autre
  $\Rightarrow$ Les relations morphologiques varient elles aussi
- Comment peut-on mesurer ces variations ?

  - En mesurant la similarité entre les mots.

- Étude réalisée sur l'anglais.
- Objectifs :
  1. mieux cerner la notion de similarité morphologique
  2. proposer des méthodes permettant d'estimer la similarité morphologique

# Types de configurations morphologiques

## Relations morphologiques paradigmatiques

| | |
|---|---|
| Famille | law:unlawful<br>admirable:admirably |
| Série | projectionist:percussionist |
| Composés | photograph:phonograph |

## Relations morphologiques non paradigmatiques

indistinguishably:readability; collectivization:decisiveness

## Absence de relation morphologique

| | |
|---|---|
| Propriétés sémantiques | legality:lawfulness; rich:wealthy |
| Propriétés phonologiques | piece:peacefully; rights:writer |
| Aucune propriété | moon:fishery |

# Contribution de l'étude

1. Faire de la similarité morphologique un objet d'étude en morphologie.
2. Proposer une articulation de la similarité morphologique avec les analyses morphématiques classiques.
3. Construire des ressources de référence destinées à la comparaison et à évaluation des mesures de la similarité morphologique.
4. Comparer 4 mesures de similarité morphologique :
    1. la distance d'édition de (Levenshtein, 1966)
    2. DPW de (De Pauw et Wagacha, 2007)
    3. Proxinette de (Hathout, 2009)
    4. Phacts de (Calderone et Celata, 2011, 2012)

# Vue d'ensemble

# Caractérisation de la similarité morphologique

- La similarité morphologique est une propriété gradable des couples de mots du lexique : 1 si identité, 0 s'il n'y a pas de relation morphologique.

**Intuition**

famille, composés > série > non paradigmatique > pas de relation

# Facteurs déterminants

**Similarité sémantique** Plus une relation morphologique connecte des mots sémantiquement similaires plus ces derniers sont morphologiquement similaires.
*acid*$_{\text{A}}$:*acidity* / *acid*$_{\text{A}}$:*acidify*

**Similarité phonologique** Plus une relation morphologique connecte des mots phonologiquement similaires plus ces derniers sont morphologiquement similaires.
*maintain*:*maintainable* / *maintain*:*maintenance*

**Complexité dérivationnelle** Plus une relation morphologique est complexe, plus les mots qu'elle réunit sont morphologiquement dissemblables.
*read*:*readable* / *read*:*unreadable*

**Canonicité** Plus une relation est régulière, plus les mots qu'elle relie sont proches.
*sweet*:*sweetness* / *long*:*length*

# Facteurs déterminants (2)

**Fréquence** La similarité entre deux mots fréquents est plus forte qu'entre deux mots rares.
*friend*:*friendship* / *ambassador*:*ambassadorship*

**Densité du réseau** Plus le réseau des relations morphologiques dans lequel se trouvent deux mots est dense, plus la relation qui les relie est forte et plus les mots sont similaires.
*familiar*:*familiarity* / *capillar*:*capillarity*

- Les facteurs sont interdépendants.
- La similarité n'est pas déterminée uniquement par les propriétés morphologiques de la relation qui s'établit entre les mots.
- Ce n'est pas une similarité exclusivement morphologique. C'est une similarité entre des mots se trouvant dans une relation morphologique.

# Vue d'ensemble

# Données

- Création de deux ressources de référence à partir de section anglaise de la base CELEX (Baayen *et al.*, 1995).
  - SMS : similarité déduite directement des représentations morphologiques de CELEX.
  - PSS : similarité estimée par la taille des paradigmes dérivationnels.

## Structure morphologique

| | |
|---|---|
| governable | `((govern)[V],(able)[A|V.])[A]` |
| traditionally | `(((tradition)[N],(al)[A|N.])[A],(ly)[B|A.])[B]` |

# Données (2)

**Lexique Large** noms, verbes, adjectifs et adverbes dont le lemme est typographiquement simple (composé uniquement de lettres minuscules).
38 670 entrées.

**Lexique Small** sous-ensemble du lexique Large réduit aux entrées dont le lemme a au moins 3 caractères et dont la fréquence est supérieure ou égale à 20.
17 887 entrées.

- L'utilisation du lexique Small est liée aux limites de la bibliothèque MaxEnt du projet OpenNLP qui a été utilisée pour calculer la mesure DPW.

# SMS: Structural matching similarity (Familles)

- Les familles réunissent les mots qui partagent une racine ou un ou plusieurs éléments de composition.
- Les représentations morphologiques de CELEX permettent de construire un graphe dérivationnel.
- La longueur du plus court chemin qui permet d'aller d'un membre d'une famille à un autre est utilisée comme estimation de leur similarité selon SMS.

**Chemin**

governable $\leftrightarrow$ govern $\leftrightarrow$ government

# SMS: Structural matching similarity (Séries)

- Les séries sont des ensembles de mots qui sont construits par la même séquence de procédés morphologiques.
  - Leurs structures morphologiques partagent une même « enveloppe externe ».
- La similarité entre une entrée et les mots de sa série peut être estimée par la complexité du schéma qui décrit la série.

## Schéma et signature

`(((navigate)[V],(able)[A|V.])[A],(ity)[N|A.])[N]`

- `((@,(able)[A|V.])[A],(ity)[N|A.])[N]`
- `(@,(ity)[N|A.])[N]`

- La similarité est estimée par le nombre de positions variables (@) et de nœuds qui portent une information catégorielle.

# SMS: Structural matching similarity (Non paradigmatiques)

- Deux mots sont NP-similaires si
  - ils n'appartiennent pas à la même famille
  - ils n'appartiennent pas à une même série
  - leur représentations contiennent au moins un affixe commun.

## Affixes communs

occupationally:industrialization

((((occupy)[V],(ation)[N|V.])[N],(al)[A|N.])[A],
(ly)[B|A.])[B]

((((industry)[N],(al)[A|N.])[A],(ize)[V|A.])[V],
(ation)[N|V.])[N]

- La similarité est estimée en comptant le nombre des affixes communs dans les représentations.

# SMS: Combiner les 3 types de similarité

- Les estimations de chaque type sont de natures totalement différentes et ne sont pas directement comparables.
- SMS ordonne les similarités conformément à l'intuition des locuteurs.
- Les membres des familles sont plus similaires les uns des autres que les membres des séries qui le sont eux-mêmes plus que les mots NP-similaires (relation non paradigmatique).

**Les 20 mots les plus similaire à *governable* selon SMS**

**govern_V** **ungovernable_A** **governance_N** **government_N** **governor_N** **misgovern_V** **governess_N** **governmental_A** **governorship_N** **guv_N** **misgovernment_N** **misgovernment_N** predictable_A comfortable_A playable_A fortifiable_A attainable_A approachable_A certifiable_A navigable_A

# SMS: Caractérisation globale

## Taille moyenne des voisinages

| Type | Large | Small |
|---|---|---|
| F | 9 | 4 |
| S | 942 | 342 |
| NP | 105 | 30 |
| SMS | 1056 | 375 |

- S $>$ SMS parce que certains mots n'ont pas de séries.
- Les familles sont des petits ensembles.
- La très grande majorité des similarités s'établissent entre les membres des séries.
- S $\gg$ NP montre qu'il existe des contraintes fortes sur les contextes dans lesquels les affixes peuvent apparaître.

# PSS: Paradigmatic strength similarity

- SMS ne permet pas de comparer les similarités de types différents.
- On peut estimer la similarité morphologique d'un couple de mots par le nombre d'analogies morphologiques auxquelles il participe.
- Analogie morphologique = analogie entre les représentations morphologiques de CELEX

## Analogies structurelles

```
(govern)[V]:((govern)[V],(able)[A|V.])[A] =
                (accept)[V]:((accept)[V],(able)[A|V.])[A]

(govern)[V]:((govern)[V],(able)[A|V.])[A] =
              (imagine)[V]:((imagine)[V],(able)[A|V.])[A]

(govern)[V]:((govern)[V],(able)[A|V.])[A] =
              (educate)[V]:((educate)[V],(able)[A|V.])[A]
```

# PSS: Paradigmatic strength similarity (2)

## Les 20 mots les plus similaires à *governable* selon PSS

**govern**_V **ungovernable**_A **government**_N **governor**_N **governance**_N manageable_A utterable_A impeachable_A endurable_A employable_A avoidable_A serviceable_A inhabitable_A favourable_A comfortable_A reliable_A **misgovernment**_N treatable_A translatable_A touchable_A

# Métriques

- Les similarités SMS et PSS ne sont définies que pour les mots de CELEX.
- Comment estimer la similarité des autres mots ?
- Comment estimer la similarités de mots d'autres langues ?

  - En calculant des similarités à partir des formes graphémiques ou des transcriptions phonologiques.

# Méthodes de base

- LCPref (*Longest Common Prefix*). La similarité des mots est estimée par la taille de leur plus long préfixe commun.
- LCSuff (*Longest Common Suffix*). La similarité des mots est estimée par la taille de leur plus long suffixe commun.

# Méthodes de base (2)

## Les 20 mots les plus similaires à *comparable*

| LCPref | LCSuff |
|---|---|
| **comparatively**_B | **incomparable**_A |
| **comparative**_A | *parable*_N |
| *compartment*_N | inseparable_A |
| **comparison**_N | unbearable_A |
| **compare**_V | bearable_A |
| *compatriot*_N | *arable*_N |
| compatible_A | arable_A |
| *compassionate*_A | vulnerable_A |
| *compassion*_N | venerable_A |
| *compass*_N | unfavourable_A |
| *company*_N | undesirable_A |
| *companionship*_N | unanswerable_A |
| companionable_A | tolerable_A |
| *companion*_N | recoverable_A |
| *compact*_N | preferable_A |
| *compact*_A | pleasurable_A |
| *computerize*_V | miserable_A |
| *computer*_N | memorable_A |
| *compute*_V | measurable_A |
| *computation*_N | invulnerable_A |

# Distance d'édition de Levenshtein

- La similarité est estimée par le nombre d'opération d'édition nécessaires pour transformer la forme d'un mot en celle d'un autre.
- opération d'édition = ajout, suppression ou remplacement d'un caractère.

- Calculs réalisés au moyen de la bibliothèque Levenshtein de Python.
- Les remplacements ont un coût de 2

$$\text{Levenshtein.ratio}(w_1, w_2) = \frac{|w_1| + |w_2| - \text{dist}(w_1, w_2)}{|w_1| + |w_2|}$$

# De Pauw and Wagacha (DPW)

- DPW permet de réaliser une analyse morphologique superficielle de langues peu dotées comme le Gĩkũyũ.
- DPW utilise un classifieur statistique basé sur le principe de l'entropie maximale pour estimer la similarité morphologique entre les mots.
- Utilisation non standard du classifieur :
  - Chaque mot du lexique définit une classe en lui-même.
  - La similarité d'un mot $y$ avec un mot $x$ est estimée par la probabilité que le mot $y$ appartienne à la classe définie par le mot $x$.
- Aucune hypothèse sur la morphologie de la langue considérée :
  - Les traits qui caractérisent les mots sont l'ensemble de $n$-grammes de lettres qui apparaissent dans leurs formes orthographiques.
  - Les $n$-grammes portent une étiquette (#) qui indique s'ils apparaissent en début, en milieu ou en fin de mot.

# De Pauw and Wagacha (DPW) (2)

## N-grammes de `comparable`

#comparable#
#comparable comparable#
#comparabl comparable omparable#
#comparab comparabl omparable mparable#
#compara comparab omparabl mparable parable#
#compar compara omparab mparabl parable arable#
#compa compar ompara mparab parabl arable rable#
#comp compa ompar mpara parab arabl rable able#
#com comp ompa mpar para arab rabl able ble#
#co com omp mpa par ara rab abl ble le#

- Calculs réalisés au moyen du système d'apprentissage automatique **csvLearner** développé par Assaf Urieli.

# Proxinette

- Proxinette a été conçue pour réduire l'espace de recherche des analogies dérivationnelles nécessaires à la découverte des paradigmes dérivationnels.
- Proxinette utilise les mêmes traits que DPW = les $n$-grammes qui apparaissent dans la forme de citation du lexème.
- Proxinette construit un bigraphe avec d'un côté les mots du lexique et de l'autre les traits qui les caractérisent.
- Chaque mot est relié à l'ensemble de ses traits.
- Chaque traits est relié à l'ensemble des mots qui le possède.

# Proxinette (2)

- Pour connaître les mots similaires à une entrée donnée,
  1. on initie une activation au niveau du sommet qui la représente.
  2. l'activation est propagée vers les traits de cette entrée.
  3. les activations qui se situent au niveau des traits sont propagées vers les mots qui les possèdent.
  4. les mots qui disposent de la plus forte activation sont ceux qui les plus similaires à l'entrée.
- Proxinette rapproche les mots qui partagent :
  - le plus grand nombre de traits communs
  - les traits les plus spécifiques = les moins fréquents.

# Phacts

- Phacts est un modèle de la formation des connaissances phonotactiques dans l'esprit des locuteurs (Calderone et Celata, 2011, 2012).
- L'algorithme est basé sur le principe des cartes auto-organisatrices de Kohonen (1995). Les cartes auto-organisatrices préservent les relations topologiques.
- Phacts utilise les mêmes traits que DPW et Proxinette = les *n*-grammes qui apparaissent dans la forme de citation du lexème.
- On projette un espace initial qui comporte autant de dimensions qu'il y a de traits utilisés pour décrire l'ensemble des mots du lexique sur une carte de $25 \times 35$ neurones = 835 dimensions.
- La similarité de deux mots est estimée par le cosinus des vecteurs qui les représentent dans la carte.

# Exemples

## Les 20 mots les plus similaires à *comparable*

| Levenshtein | DPW | Proxinette | Phacts |
|---|---|---|---|
| **incomparable**_A | **incomparable**_A | **incomparable**_A | honourable_A |
| *parable*_N | *parable*_N | **incomparably**_B | comfortable_A |
| **compare**_V | arable_A | **comparatively**_B | conceivable_A |
| **incomparably**_B | *arable*_N | **comparative**_A | commendable_A |
| compatible_A | inseparable_A | *parable*_N | hospitable_A |
| companionable_A | **compare**_V | inseparable_A | formidable_A |
| **comparative**_A | unbearable_A | *compartment*_N | considerable_A |
| comfortable_A | companionable_A | **comparison**_N | charitable_A |
| *marble*_N | compatible_A | **compare**_V | noticeable_A |
| *arable*_N | *company*_N | unbearable_A | measurable_A |
| arable_A | bearable_A | bearable_A | creditable_A |
| incompatible_A | durable_A | *arable*_N | contemptible_A |
| payable_A | *compact*_A | arable_A | deplorable_A |
| *comrade*_N | adorable_A | *parabolic*_A | remarkable_A |
| *complex*_N | vulnerable_A | *unbearably*_B | *monosyllable*_N |
| *complex*_A | inexorable_A | companionable_A | favourable_A |
| *compile*_V | **comparison**_N | compatible_A | answerable_A |
| capable_A | tolerable_A | *compatriot*_N | marketable_A |
| remarkable_A | **comparative**_A | *compassionate*_A | preferable_A |
| measurable_A | *compact*_N | *compassion*_N | foreseeable_A |

# Évaluation des voisinages morphologiques

## Précision, Rappel, Fscore

- Ces mesures permettent de vérifier la capacité des métriques à capter les trois types de similarité
  1. dans les familles dérivationnelles,
  2. dans les séries dérivationnelles,
  3. entre les mots qui sont dans des relations non paradigmatiques

$$P = \frac{|V \cap S|}{|V|} \qquad R = \frac{|V \cap S|}{|S|} \qquad F = \frac{2\ P\ R}{P + R}$$

- $V$: ensemble des voisins relativement à la métrique candidate
- $S$: ensemble des voisins relativement à la métrique de référence

# Évaluation des voisinages morphologiques (2)

## $\tau$ de Kendall

- Cette mesure permet de vérifier la capacité des métriques à ordonner les voisins selon un ordre compatible avec l'intuition :
  - similarité plus forte dans les familles
  - similarité plus faible dans les séries
  - similarité négligeable entre les mots qui sont dans des relations non paradigmatiques

$$inv_{XY}(x,y) = \begin{cases} 1 & \text{if } x \in X,\ y \in Y \text{ and } r(x) > r(y) \\ 0 & \text{otherwise} \end{cases}$$

$$\tau = \frac{\sum_{(x,y)\in X\times Y} inv_{XY}(x,y)}{|X \times Y|}$$

- $r(x)$ = rang de $x$ dans la liste des voisins candidats.

# Évaluation des voisinages morphologiques (3)

## P@N

- Précision au rang $N = 1, 2, 5, 10, 20$ et 100.
- Cette mesure compare les métriques à la référence PSS pour déterminer leur capacité globale à ordonner correctement l'ensemble des voisins, indépendamment des types de similarité.

# Vue d'ensemble

# Résultats

## Taille moyenne des vosinages

| | Large | Small |
|---|---|---|
| SMS | 1055 | 375 |
| PSS | 327 | 66 |
| LCPref | 440 | 214 |
| LCSuff | 1003 | 382 |
| Levenshtein | 569 | 309 |
| DPW | – | 300 |
| Proxinette | 488 | 276 |
| Phacts | 500 | 300 |

- SMS $\gg$ PSS
- La majorité des relations dérivationnelles ne sont pas régulières.
- 1/3 à 1/5 seulement d'entre elles forment des analogies.
- La proportion de relations régulières augmente avec la taille du corpus.

# Résultats (2)

## P, R, F relativement aux familles de SMS

| | Large | | | Small | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| PSS | 0.013 | 0.541 | 0.026 | 0.025 | 0.589 | 0.049 |
| LCPref | 0.011 | 0.423 | 0.022 | 0.016 | 0.532 | 0.031 |
| LCSuff | 0.003 | 0.284 | 0.006 | 0.003 | 0.101 | 0.005 |
| Levenshtein | 0.015 | 0.694 | 0.030 | 0.019 | 0.820 | 0.036 |
| DPW | – | – | – | **0.023** | 0.928 | **0.044** |
| Proxinette | **0.023** | **0.948** | **0.045** | 0.022 | **0.972** | 0.043 |
| Phacts | 0.004 | 0.154 | 0.007 | 0.005 | 0.217 | 0.009 |

- Les familles sont de petits ensembles : précision faible / rappel fort.
- DPW et Proxinette se détachent.

# Résultats (3)

## P, R, F relativement aux séries de SMS

| | Large | | | Small | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| PSS | 0.982 | 0.422 | 0.590 | 0.981 | 0.314 | 0.476 |
| LCPref | 0.074 | 0.017 | 0.028 | 0.081 | 0.019 | 0.031 |
| LCSuff | **0.512** | **0.363** | **0.425** | **0.439** | **0.108** | **0.173** |
| Levenshtein | 0.291 | 0.087 | 0.134 | 0.237 | 0.075 | 0.114 |
| DPW | – | – | – | 0.354 | 0.107 | 0.164 |
| Proxinette | 0.227 | 0.062 | 0.098 | 0.203 | 0.069 | 0.103 |
| Phacts | 0.283 | 0.082 | 0.128 | 0.274 | 0.098 | 0.144 |

- Les séries sont de gros ensembles : précision forte / rappel faible.
- La **méthode de base** LCSuff dépasse toutes les autres.
- DPW obtient de bons résultats

# Résultats (4)

## P, R, F relativement aux NP-similaires de SMS

| | Large | | | Small | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| PSS | 0.013 | 0.042 | 0.020 | 0.016 | 0.021 | 0.018 |
| LCPref | 0.072 | 0.096 | 0.082 | 0.085 | 0.105 | 0.094 |
| LCSuff | 0.031 | **0.153** | 0.051 | 0.034 | 0.055 | 0.042 |
| Levenshtein | 0.060 | 0.107 | 0.077 | 0.077 | 0.125 | 0.095 |
| DPW | – | – | – | 0.085 | 0.126 | 0.101 |
| Proxinette | **0.074** | 0.119 | **0.091** | **0.092** | **0.168** | **0.119** |
| Phacts | 0.042 | 0.077 | 0.054 | 0.059 | 0.119 | 0.079 |

- Résultats difficilement interprétables.

# Résultats (5)

## P, R, F relativement à SMS

| | Large | | | Small | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| PSS | 1.000 | 0.376 | 0.547 | 1.000 | 0.282 | 0.439 |
| LCPref | 0.085 | 0.028 | 0.043 | 0.083 | 0.030 | 0.044 |
| LCSuff | **0.416** | **0.341** | **0.375** | **0.287** | 0.102 | 0.151 |
| Levenshtein | 0.221 | 0.093 | 0.131 | 0.172 | 0.085 | 0.114 |
| DPW | – | – | – | 0.250 | **0.114** | **0.157** |
| Proxinette | 0.201 | 0.073 | 0.107 | 0.171 | 0.083 | 0.112 |
| Phacts | 0.218 | 0.082 | 0.119 | 0.202 | 0.100 | 0.134 |

- LCSuff est la meilleure sur les séries
- Plus de 90% des similarités s'établissent dans les séries.

# Résultats (6)

## Proportion des inversions relativement à SMS

| | Large | | | Small | | |
|---|---|---|---|---|---|---|
| | F×S | F×NP | S×NP | F×S | F×NP | S×NP |
| PSS | 0.076 | 0.104 | 0.449 | 0.047 | 0.061 | 0.560 |
| LCPref | **0.119** | **0.126** | 0.483 | **0.099** | 0.103 | 0.499 |
| LCsuff | 0.323 | 0.175 | 0.452 | 0.229 | **0.101** | 0.517 |
| Levenshtein | 0.257 | 0.208 | 0.410 | 0.208 | 0.170 | 0.401 |
| DPW | – | – | – | 0.217 | 0.190 | 0.397 |
| Proxinette | 0.211 | 0.202 | 0.463 | 0.129 | 0.120 | 0.472 |
| Phacts | 0.367 | 0.296 | **0.332** | 0.396 | 0.309 | **0.389** |

# Résultats (7)

- LCPref ne ramène que des membres de la famille $\rightarrow$ pas d'inversions
- Les trois types se répartissent en 2 groupes :
  1. les familles (partage d'un radical)
  2. les séries et les NP-similaires (partage d'un ou de plusieurs affixes)
- Les *n*-grammes de caractères ne permettent pas de séparer les séries des mots NP-similaires.

# Résultats (8)

## Précision à $N$ relativement à PSS

Large

| | P@1 | P@2 | P@5 | P@10 | P@20 | P@50 | P@100 |
|---|---|---|---|---|---|---|---|
| LCPref | **0.448** | **0.518** | 0.441 | 0.331 | 0.231 | 0.144 | 0.106 |
| LCSuff | 0.064 | 0.110 | 0.168 | 0.220 | 0.265 | **0.351** | **0.440** |
| Levenshtein | 0.345 | 0.373 | 0.357 | 0.326 | 0.300 | 0.310 | 0.344 |
| Proxinette | 0.310 | 0.459 | **0.470** | **0.414** | **0.347** | 0.311 | 0.316 |
| Phacts | 0.015 | 0.027 | 0.068 | 0.117 | 0.158 | 0.231 | 0.306 |

Small

| | P@1 | P@2 | P@5 | P@10 | P@20 | P@50 | P@100 |
|---|---|---|---|---|---|---|---|
| LCPref | **0.511** | **0.550** | 0.429 | 0.302 | 0.206 | 0.132 | 0.104 |
| LCSuff | 0.035 | 0.063 | 0.101 | 0.141 | 0.184 | 0.257 | 0.323 |
| Levenshtein | 0.383 | 0.425 | 0.421 | 0.378 | 0.364 | 0.368 | 0.375 |
| DPW | 0.229 | 0.323 | 0.399 | 0.410 | **0.436** | **0.485** | **0.518** |
| Proxinette | 0.375 | 0.514 | **0.518** | **0.434** | 0.373 | 0.348 | 0.363 |
| Phacts | 0.019 | 0.045 | 0.135 | 0.192 | 0.261 | 0.346 | 0.418 |

# Résultats (9)

- Les deux premiers voisins sont des membres de la famille $\rightarrow$ LCPref ramène les membres de la famille qui sont les plus proches (dérivations au moyen de suffixes courts).
- Proxinette ramène une bonne partie de la famille et les membres de la série qui sont les plus proches.

Large LCSuff l'emporte au delà durant 50

Small DPW a les meilleures performances globales au delà du rang 20

# Vue d'ensemble

# Conclusion

- Identifier les mots similaires apparaît comme une tâche

  simple des heuristiques élémentaires comme LCPref pour les familles et LCSuff pour les séries obtiennent des résultats qui se situent parmi les meilleurs.

  complexe aucune des mesures considérées n'est capable de capter la similarité morphologique dans sa globalité.

- Les performances globales des métriques sont moyennes. Les résultats ne pourront être améliorés que par une réelle prise en compte du sens. Cependant :
  - les MRD sont difficiles à exploiter
  - les bases distributionnelles sont trop bruitées
  - les bases lexicales comme WordNet sont trop spécialisées.

# Conclusion (2)

- Les meilleurs résultats sont obtenus par la méthode par apprentissage statistique DPW.
  - Les régularités morphologiques doivent être considérées au niveau du lexique dans sa globalité
- Les similarités non paradigmatiques sont marginales relativement aux similarités paradigmatiques.
  - Elles ne se distinguent pas fortement des similarités entre les membres des séries.
  - Les NP-similarités sont des relations accidentelles qui n'interviennent pas dans la structuration du lexique.

# Perspectives

- Étudier et comparer la validité psychologique de la similarité morphologique, des facteurs qui la déterminent et des métriques qui permettent de l'estimer.
- Adapter ce travail aux relations morphologiques dérivationnelles.
  - Il n'existe pas de référentiel.
  - Comment pondérer les différents traits flexionnels : genre, nombre, personne, temps, mode, etc. ?
  - 2 formes verbales qui partagent le même temps sont-elles plus similaires que 2 formes verbales qui partagent la même personne ?

# Références


Baayen, R. H., Piepenbrock, R. et Gulikers, L. (1995). The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, Philadelphia, PA.

Calderone, B. et Celata, C. (2011). Paradigm-aware morphological categorizations. *Lingue e linguaggio*, 2011(2):183–207.

Calderone, B. et Celata, C. (2012). PHACTS about activation-based word similarity effects. *In Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss*, pages 33–37, Avignon. ACL.

De Pauw, G. et Wagacha, P. W. (2007). Bootstrapping morphological analysis of Gĩkũyũ using Maximum Entropy Learning. *In Proceedings of the Eighth Annual Conference of the International Speech Communication Association (Interspeech)*, Antwerp, Belgique.

# Références (2)

Hathout, N. (2009). Acquisition of morphological families and derivational series from a machine readable dictionary. *In* Montermini, F., Boyé, G. et Tseng, J., éditeurs : *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*, Somerville, MA. Cascadilla Proceedings Project.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer Verlag, Berlin / Heidelberg.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10(8):707–710.