Mesurer la distance morphologique

Nabil HATHOUT, Basilio CALDERONE, Fabio MONTERMINI

CLLE-ERSS, CNRS & UTM

Séminaire de l'UE TAL

12 décembre 2011

Introduction

- Qu'est ce qu'une mesure de la distance morphologique ?
- À quoi ça peut servir ?
- Comment ça peut se calculer ?
- Comment comparer / évaluer ces mesures ?

Distance morphologique

- Une mesure de la distance morphologique est une fonction qui associe à chaque couple de mots du lexique une valeur numérique.
- Deux mots sont en relation morphologique s'ils partagent en même temps une partie de leur forme et de leur sens.

- Il ne s'agit pas de distance au sens mathématique du terme (symétrie, séparation, inégalité triangulaire).
- Si w_1 et w_2 sont en relation morphologique, $D(w_1, w_2) > 0$
- Si w_1 et w_2 n'ont aucune relation morphologique, $D(w_1, w_2) = 0$
- Si w_1 est construit à partir de w_2 et w_2 est construit à partie de w_3, alors $D(w_1, w_2) < D(w_1, w_3)$

$D(\text{incorrection\_N,incorrect\_A}) < D(\text{incorrection\_N,correct\_A})$

Vue d'ensemble

Aspects linguistiques

- Le lexique est une entité structurée.
- Chaque élément lexical est inséré dans un environnement (un voisinage).
- Par exemple, il fait partie de familles et de séries morphologiques.
- Il est possible de quantifier la ressemblance des membres de ces ensembles.

Aspects linguistiques

- En théorie, il serait possible de mesurer la relation de n'importe quelle unité du lexique avec n'importe quelle autre unité.
- En pratique, seules certaines relations sont pertinentes / intéressantes.
  - *plage* / *sable*
  - *admirable* / *sable*
  - *admirable* / *éligible*
  - *admirable* / *admiration*
  - *éligible* / *élection*

Aspects linguistiques

Distance formelle (phonologique) : en fonction des segments (phonèmes) partagés et de leur distribution.

Distance sémantique : en fonction des traits sémantiques partagés, des propriétés distributionnelles.

Distance morphologique : en fonction de la densité de l'environnement lexical, du nombre de voisins (des membres de la famille ou des sous-séries) partagés.

Aspects linguistiques

- Dans plusieurs domaines (phonologie, morphologie, syntaxe) on se sert de distinctions qui font référence à la distance entre deux mots :
  - allomorpie vs. supplétion
  - transparent vs. non transparent
- En général, on fait comme s'il s'agissait de distinctions qualitatives et qu'il était possible d'étendre à tout le lexique ce qu'on observe localement pour des couples de mots.

Aspects linguistiques

- Donner une métrique quantitative à des différences supposées qualitatives.
- On peut continuer de se servir des mêmes catégories mais en se basant sur des mesures objectives.

Aspects linguistiques

Quelques applications :

- Linguistique théorique (phonologie, morphologie, syntaxe, etc.) :
  - allomorphie, transparence, cas de déviation forme / sens (*anticancéreux*)
- Psycholinguistique :
  - transparence, taille / densité de la famille lexicale
- TAL :
  - les voisinages permettent de tenir compte des variations morphologiques dans les systèmes de RI ;
  - dégrossissage des relations morphologiques dans un lexique ;
  - réduire l'espace de recherche lorsqu'on cherche à connecter les mots qui sont morphologiquement apparentés.

Vue d'ensemble

CELEX

- Comparer différentes mesures qui permettent d'**estimer la parenté morphologique** qui existe entre les mots du dictionnaire.
- La comparaison est réalisée relativement à un lexique de référence : la partie anglaise de la base CELEX.
- CELEX est la seule base de données morphologiques dérivationnelles largement diffusée et utilisée (anglais, allemand, néerlandais).

CELEX

- CELEX contient des informations
  - phonologiques et graphémiques ;
  - **morphologiques** (flexionnelles et **dérivationnelles**) ;
  - catégorielles ;
  - fréquences par lemme et par forme.

Quelques informations fournies par CELEX

acoustically/B/6/C/SA/**((acoustic)[A],(ally)[B|A.])[B]**/derivation

applicable/A/81/C/SA/**((apply)[V],(able)[A|V.])[A]**/derivation

applique/N/1/Z/S/**((applique)[V])[N]**/conversion

applejack/N/1/C/SS/**((apple)[N],(jack)[N])[N]**/composition

aerodrome/N/10/C/AA/**((aero)[N|.x],(drome)[N|x.])[N]**/neoclassique

- Comment calculer des distances entre les mots à partir de leurs **décompositions morphématiques** ?

Graphe CELEX

- Les décompositions morphématiques sont analysées en utilisant Lex et Yacc.
- Construction d'un graphe CELEX
  - Les sommets sont les décompositions des lexèmes et des élements de composition.
  - Les arcs relient les décompositions des lexèmes complexes aux décompositions de leurs constituants. Elles sont symétriques.

Exemples d'arcs dans le graphe CELEX

acoustically_B $\leftrightarrow$ **acoustic_A**

apologize_V $\leftrightarrow$ **apology_N**

applicable_A $\leftrightarrow$ **apply_V**

applique_N $\leftrightarrow$ **applique_V**

applejack_N $\leftrightarrow$ **apple_N** ; **applejack_N** $\leftrightarrow$ **jack_N**

aerodrome_N $\leftrightarrow$ **aero_N.x** ; **aerodrome_N** $\leftrightarrow$ **drome_N.x**

Familles dérivationnelles dans CELEX

- Le graphe est parcouru pour constituer des familles morphologiques.

1. On calcule la fermeture symétrique et transitive des relations de dérivations et de conversion.
2. Pas de fermeture transitive (relations transversales) pour les composés.
   - Un mot composé est dans les familles dérivationnelles de tous ses éléments de composition.
   - Un composant appartient à la famille dérivationnelle de chaque composé dont il est l'un des éléments.

aerodrome et _airplane_ n'appartiennent pas à la même famille

aeroplane_N $\Leftrightarrow$ aerodrome_N, airplane_N

airplane_N $\not\Leftrightarrow$ aerodrome_N

aerodrome_N $\leftrightarrow$ aero_N.x ; aerodrome_N $\leftrightarrow$ drome_N.x

aeroplane_N $\leftrightarrow$ aero_N.x ; aeroplane_N $\leftrightarrow$ plane_N

airplane_N $\leftrightarrow$ air_N ; airplane_N $\leftrightarrow$ plane_N

Familles dérivationnelles dans CELEX

Distance simple entre les membres d'une famille

Si w_1 et w_2 appartiennent à la même famille, on peut définir $D_f(w_1, w_2)$ comme le nombre d'arcs du plus court chemin entre w_1 et w_2.

Distance des éléments de la famille de *determinant_N*

determinable_A=2, determinant_A=2, determinant_N=0, determinate_A=2, determination_N=2, determinative_A=3, determinative_N=4, determine_V=1, determiner_N=2, determinism_N=2, deterministic_A=3, indeterminable_A=3, indeterminably_B=4, indeterminacy_N=4, indeterminate_A=3, predetermination_N=3, predetermination_N=3, predetermine_V=2, predeterminer_N=3

Distance des éléments de la famille de *fruit_N*

breadfruit_N=1, fruit_N=0, fruit_V=1, fruitcake_N=1, fruiterer_N=1, fruitful_A=1, fruitfully_B=2, fruitfulness_N=2, fruition_N=1, fruitless_A=1, fruitlessly_B=2, fruitlessness_N=2, fruity_A=1, grapefruit_N=1, unfruitful_A=2

Séries dérivationnelles dans CELEX

- Un mot appartient généralement à plusieurs sous-séries.

_indeterminably_B_ appartient aux sous-séries des adverbes en :

in-V-ably : incurably_B, indisputably_B, inexplanably_B

_in-A-ly_ : inhumanely_B, incautiously_B, incomptently_B

_A-ly_ : disapprouvingly_B, sincerely_B, categorically_B

- Les sous-séries d'une entrée peuvent être calculées à partir de sa décomposition en réalisant des **abstractions** sur ses constituants.

Signatures des sous-séries de _indeterminably_B_

décomposition = (((in)[A|.A],((determine)[V],(able)[A|V.])[A])[A],(ly)[B|A.])[B]

-

- (((in)[A|.A],(@,(able)[A|V.])[A])[A],(ly)[B|A.])[B]
- (((in)[A|.A],@)[A],(ly)[B|A.])[B]
- (@,(ly)[B|A.])[B]

Séries dérivationnelles dans CELEX

Distance simple entre un mot et les éléments de ses sous-séries

1. Ordonner les sous-séries de w par ordre décroissant du nombre de symboles de leurs signatures.
2. Si σ est une sous-séries de w, $D_s(w, \sigma)$ est le rang de σ parmi les sous-série de w.
3. Si w' appartient aux sous-séries $\sigma_1, \cdots, \sigma_n$ de w,

$$D_s(w, w') = \min_{i=1 \cdots n} D_s(w, \sigma_i)$$

Comment combiner les mesures de distance morphologiques dans les familles et les sous-séries ?

Mesure de référence

- Proposer une mesure unique pour les membres des familles et des sous-séries morphologiques.
- L'idée est d'estimer la distance morphologique entre deux mots en fonction du nombre d'analogies dans lesquels ils apparaissent :

Distance analogique

1. Rechercher toutes les analogies morphologiques qui s'établissent entre les mots du lexique.
2. Définir une distance D_r telle que

$$D_r(w_1, w_2) = \frac{|\{(x, y) \in L \times L / w_1 : w_2 = x : y\}|}{N(w_1)}$$

   - où L est l'ensemble des mots du lexique
   - $a : b = c : d$ est vrai s'il existe une analogie formelle entre les décompositions morphématiques de a, b, c et d
   - $N(w)$ est le nombre total d'analogies dans lesquelles apparaît w.

Analogies morphologiques structurelles

academic :academically=psychedelic :psychedelically

ϵ	((academy)[N],(ic)[A|N.])[A]	ϵ
(	((academy)[N],(ic)[A|N.])[A]	,(ally)[B|A.])[B]

ϵ	(psychedelic)[A]	ϵ
(	(psychedelic)[A]	,(ally)[B|A.])[B]

- Calculer toutes les analogies qui s'établissent entre toutes les décompositions de CELEX est difficile à réaliser car de trop grande complexité computationnelle.

Analogies morphologiques structurelles

- On peut collecter un grand nombre de ces analogies en associant à chaque couple une signature qui décrit les opérations d'édition qui permettent de passer d'une décomposition à l'autre.

academically :*academic*

```
D/['('']/[] ; D/['(ally)', '[B|A.]', ')', [B]']/[]
```

- On utilise la `difflib` (l'algorithme du `diff` de Unix) en faisant varier le découpage des décompositions.
- Nous avons calculé 31 millions d'analogies pour les 41 120 entrées de CELEX.
- Les analogies dont l'une des relations est une conversion sont exclues.

Analogies morphologiques structurelles

20 voisins les plus proches de *avoidable_A*

avoid_V :0.445826 unavoidable_A :0.422735 unavoidably_B :0.113677
avoidance_N :0.017762 favourable_A :0.008881 comfortable_A :0.008881
utterable_A :0.007105 speakable_A :0.005329 serviceable_A :0.005329 readable_A :0.005329
questionable_A :0.005329 mistakable_A :0.005329 inhabitable_A :0.005329
impeachable_A :0.005329 governable_A :0.005329 forgettable_A :0.005329
fathomable_A :0.005329 endurable_A :0.005329 desirable_A :0.005329 deniable_A :0.005329
believable_A :0.005329 bearable_A :0.005329 accountable_A :0.005329

- L'évaluation des mesures est faite en termes de Précision, Rappel et Fscore pour les tailles de voisinages de 5, 10, 20, 50, 100, 200, 500 et 1000 voisins. Le Fscore est la moyenne harmonique de la Précision et du Rappel.

Vue d'ensemble

Autonomie de la morphologie

Autonomie de la morphologie

Dans les présentes simulations, l'effort de quantifier une mesure de 'morphologicité' est basé exclusivement sur les aspects de 'surface' des mots, sans l'adjonction d'informations sur la composante sémantico-contextuelle des mots (*blind-to-semantics framework*)

Que reste sans la sémantique ?

- Analogies et relations purement formelles entre les mots
- Effets de fréquence (type et token)
- Aspects phonotactiques de la langue
- Structures paradigmatiques

Autonomie de la morphologie

Motivations

- **Evidences comportementales et psycholinguistiques** :
  - Existence de effets morphologiques en 'word recognition' qui ne sont pas influencés par la transparence sémantique. (early visual recognition : Taft, 1994 ; Taft, 1979 ; Caramazza et al., 1988 ; Schreuder and Baayen, 1995 ; Meunier and Segui, 2002 ; Rastle and Davis, 2004)
  - Acquisition de la morphologie et maîtrise de la competence morphologique (Pinker et Ullman, 2002 ; Pirrelli et al. 2011)

- **Evidences linguistiques-descriptives** :
  - Structure et organization du lexique : analyses de type 'stem space' (Bonami et Boyé, 2003 ; Boyé et Montermini, 2011)
  - Structures paradigmatiques emergents (Giraudo et al. in press ; Albright, 2009 ; Calderone et Celata, 2011)

Autonomie de la morphologie

Attention

En affermant l'autonomie de la morphologie pour certains structures/niveaux linguistiques, nous ne sommes pas en train de caractériser la morphologie en termes de modularité

Vue d'ensemble

Distance de Levenshtein

Notion basique

La **distance de Levenshtein** mesure la similarité entre deux chaînes (sequences) S et T de caractères. Elle est égale au nombre minimal de caractères qu'il faut *supprimer*, *insérer* ou *remplacer* pour passer d'une chaîne à l'autre.

- Mesure de surface $\rightarrow$ *matching* de caractères (des techniques d'alignement)
- Plusieurs versions : normalisation par nb. de caractères, différents coûts de passage, etc.

Distance de Levenshtein

Matrice de Levenshtein

- Si $|S| = m$et $|T| = n$ il suffit construire une matrice de taille $m + 1 * n + 1$
- $D(0,0) = 0$, $D(i,0) = i$ où $i = 1 \ldots m$, $D(0,j) = j$ où $j = 1 \ldots n$
- Remplir la matrice avec une itération des coûts :
  - Si $S(i) == T(j)$ coût $= 0$, si no coût$= 1$
  - $D(i,j) = min(D(i-1,j-1)+\text{coût}, D(i,j-1)+1, D(i-1,j)+1)$

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

Distance de Levenshtein

Voisins pour *direction_N*

directions_N=0.947368421053, ***directional_A***=0.9, diction_N=0.875, misdirection_N=0.857142857143, dissection_N=0.842105263158, discretion_N=0.842105263158, reaction_N=0.823529411765, erection_N=0.823529411765, ***director_N***=0.823529411765, imprecation_N=0.8, distraction_N=0.8, ***direct_V***=0.8, ***direct_B***=0.8, ***direct_A***=0.8, diffraction_N=0.8, dereliction_N=0.8, deprecation_N=0.8, rejection_N=0.777777777778, reduction_N=0.777777777778, redaction_N=0.777777777778, reception_N=0.777777777778, injection_N=0.777777777778, infection_N=0.777777777778, ***directory_N***=0.777777777778, ***directive_N***=0.777777777778, directive_A=0.777777777778, digestion_N=0.777777777778, dictation_N=0.777777777778, detection_N=0.777777777778, depiction_N=0.777777777778, dejection_N=0.777777777778, defection_N=0.777777777778, deduction_N=0.777777777778, deception_N=0.777777777778, bisection_N=0.777777777778, addiction_N=0.777777777778, intersection_N=0.761904761905, interjection_N=0.761904761905, insurrection_N=0.761904761905, indiscretion_N=0.761904761905, imperfection_N=0.761904761905, disinfection_N=0.761904761905

Distance de Levenshtein

Réflexion

- Il s'agit d'une distance très bonne pour trouver les mots qui partagent un maximum de séquences de lettres communes
- Mais elle ne prend pas en compte la position du caractère modifié dans la chaîne : le fait qu'une lettre soit ajoutée/supprimée/modifiée au début, au milieu ou à la fin du mot n'a pas d'importance.

- Pour la presente simulation :
  - Calcul de la distance de Levenshtein pour les 30.000 mots de CELEX
  - Repérage de *n*-voisins pour chaque mot de CELEX

Vue d'ensemble

PHACTS : un système basé sur la phonotaxe

Phonotaxe

- C'est quoi la phonotaxe d'une langue ?
- À quoi ça peut servir ?
- Comment peut-on calculer une matrice phonotactique ?
- Comment peut-on 'construire' une représentation de mot basée sur la phonotaxe ?

Phonotaxe

- C'est quoi la phonotaxe d'une langue ?
  - Combinaisons de phonèmes (ou caractères) attestés et récurrents dans la langue

En français

Le phonème /*o*/ peut être transcrit par plusieurs séquences de graphèmes : $< o >, \ldots < au >, \ldots < eau >, \ldots < ot >$ *etc*

Souvent le graphème utilisé dépend de plusieurs variables :

- Position de /o/ dans le mot
- Son contexte consonantique

Par exemple, en position centrale, /*o*/ est plus souvent transcrit comme $< o >$ que $< au >$ entre $< b >$ et $< r >$, mais est bien souvent transcrit comme $< au >$ (en comparaison à $< o >$) entre $< p >$ et $< v >$. En position final /*o*/ est fréquemment transcrit $< eau >$ après $< r >$ ou $< t >$, mais il n'est jamais transcrit comme $< eau >$ après $< f >$.

Construire mots à partir de la phonotaxe d'une langue

- Comment peut-on exploiter la connaissance phonotactique pour créer une représentation (psycholinguistiquement légitime) du mot (conçu comme une séquence temporelle d'unites phonologiques) ?
- Comment peut-on intégrer dans cette représentation les effets distributionnels de fréquence des séquences fortement attestées (en termes de *type* et *token*) ?

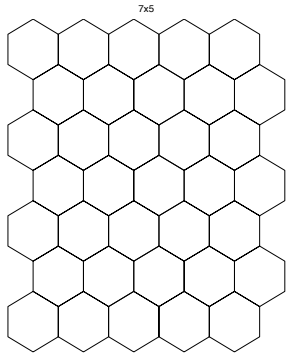
À la base de PHACTS : cartes adaptatives

Inspiration : cartes auto-organisatrices (Kohonen 2001), mémoires associatives (Hopfield 1982)

- Systèmes probabilistes d'organisation des données d'input
- Structure physique (en théorie) et computation en parallèle

À la base de PHACTS : cartes adaptatives

- D'un point de vue architectural, les cartes adaptatives sont constituées d'une grille de noeuds ou neurones (le plus souvent uni- ou bidimensionnelle)
- Dans chaque noeud de la grille se trouve un neurone-vecteur (composants random au début, avant apprentissage)
- Chaque neurone est (en théorie) lié à un vecteur référent, responsable d'une position dans l'espace des données d'input
- L'idée de base est d'associer chaque donnée d'input à un (ou plusieurs) neurone sur la carte (après apprentissage)

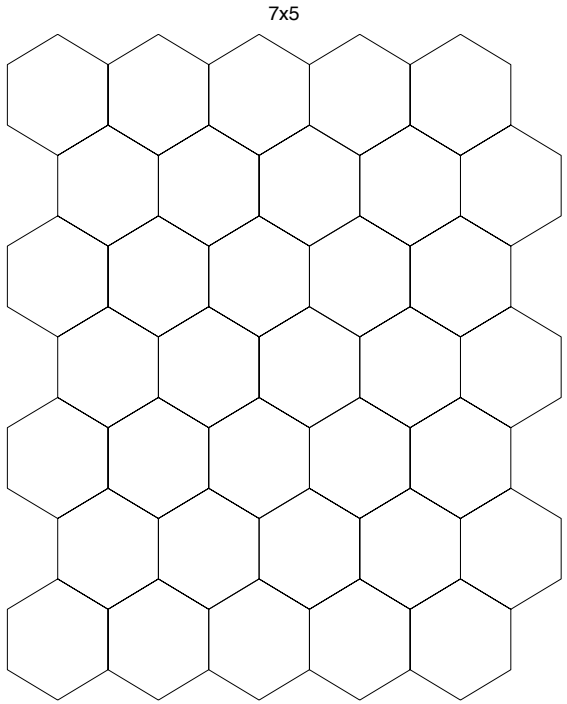


NB : Les données structurellement similaires doivent être représentées par des neurones adjacents sur la carte $\Rightarrow$ *Clustering*

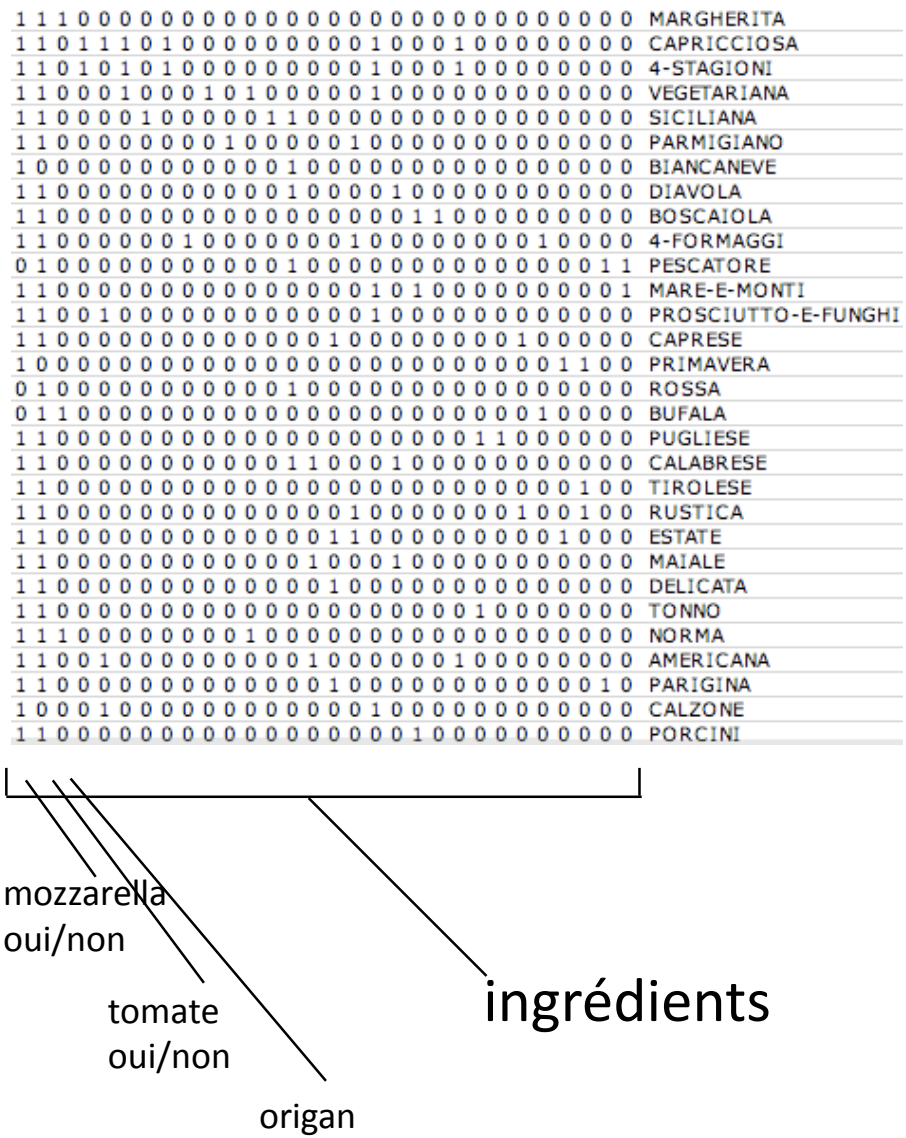
À la base de PHACTS : cartes adaptatives

Questions

- Mais quel type des données ?
- Mais quel type de clusters ?
- Mais quel type d'apprentissage ?
- Où est la phonotaxe ?



Cartes adaptatives : un exemple italien



- Données : vecteur de pizzas
- Chaque composant définit la présence/absence d'un ingrédient

Cartes adaptatives : un exemple italien

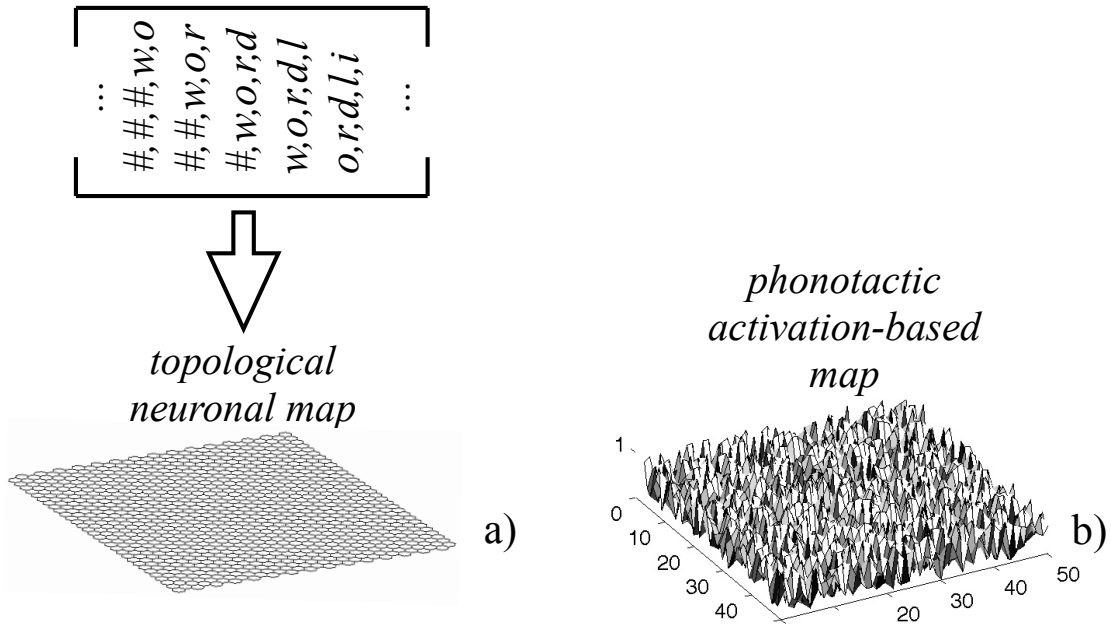
- Organisation topologique des pizzas
- En bas : pizzas *fromageuses*
  Au milieu : pizzas *charcuteries*
  En haut (gauche) : pizzas végétariennes
- La carte adaptative se déploie de façon à représenter un ensemble de données, et chaque neurone se spécialise pour représenter un groupe particulier des données selon les points communs qui les rassemblent
- Techniquement, la carte réalise une quantification vectorielle de l'espace des données

PHACTS : l'architecture

Idée : Obtenir une organisation topologique de la phonotaxe de la langue

- Pas ingrédients, mais *n-grams* transcrits de manière binaire (orthogonale) : chaque caractère est indépendant des autres
- Données : 30.000 mots de CELEX
- Dimension de la carte : 100*x*110
- Métrique d'apprentissage (vecteur-input vs vecteur-référent) : inner product

PHACTS : l'architecture



- Importance de l'activation de chaque neurone : l'activation est liée à la fréquence de token du *n-gram* considéré
- Importance de la position de chaque neurone : la position est liée au type de *n-gram* considéré
- Importance de la position du *n-gram* dans les mots : les *n-grams* 'right-side' (fin de mot) sont les plus actifs

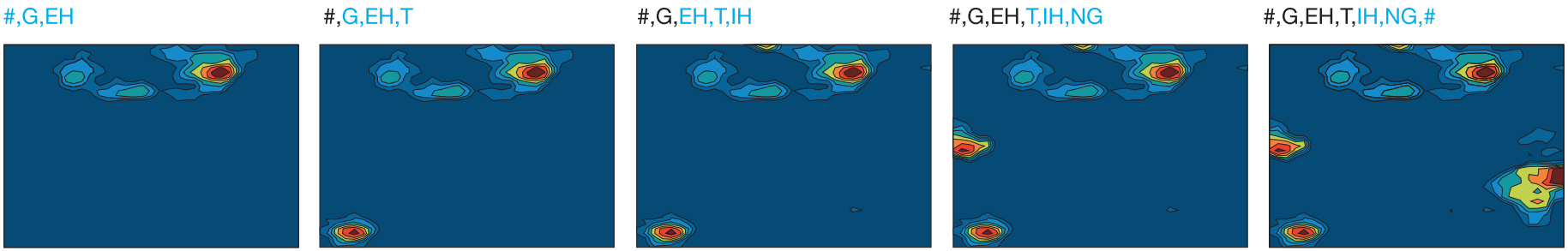
PHACTS : l'architecture

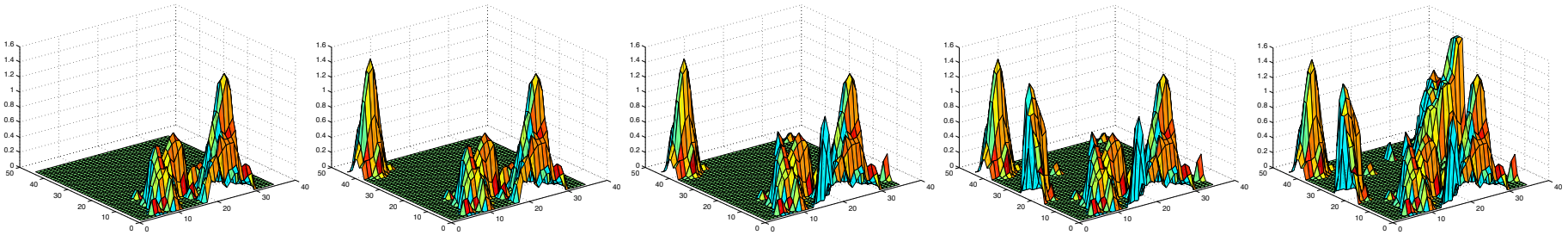
Idée : Reconstruire le mot à partir de cette carte phonotactique

- Reconstruire une dimensione de mot en échantillonnant tous ses *n-grams* au moyen de la carte phonotactique
- La sommation des états d'activation déclenchés par des *n-grams* définit une représentation vectorielle où les mots sont recodées sur la base de la connaissance phonotactique préalablement acquise

$F_{\text{phacts}}(mot) = \sum \Phi(ngrams)$
Φ =activations de la carte phonotactique

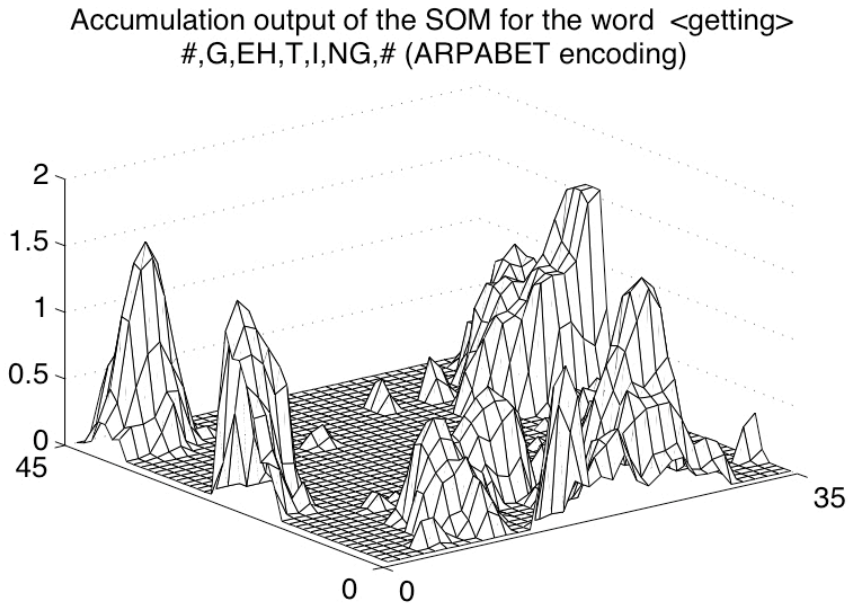
PHACTS : l'architecture

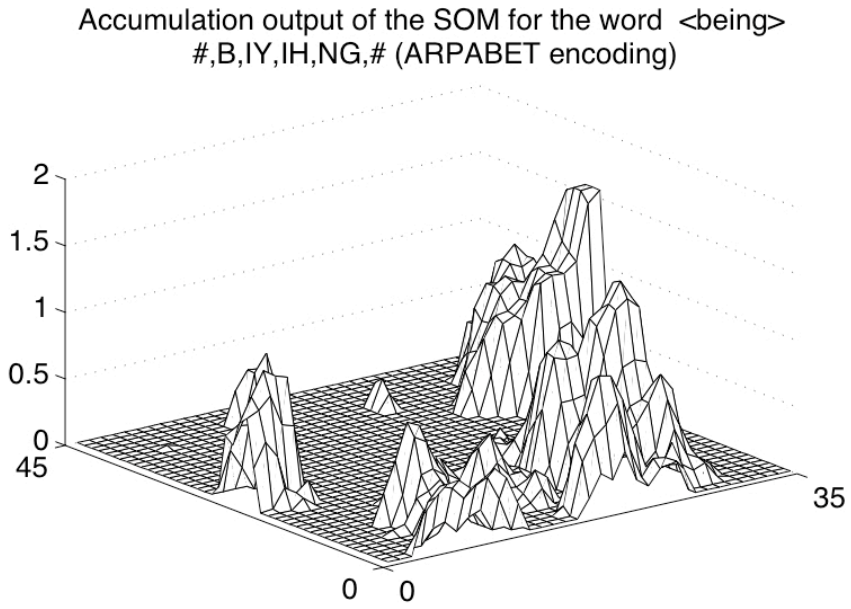


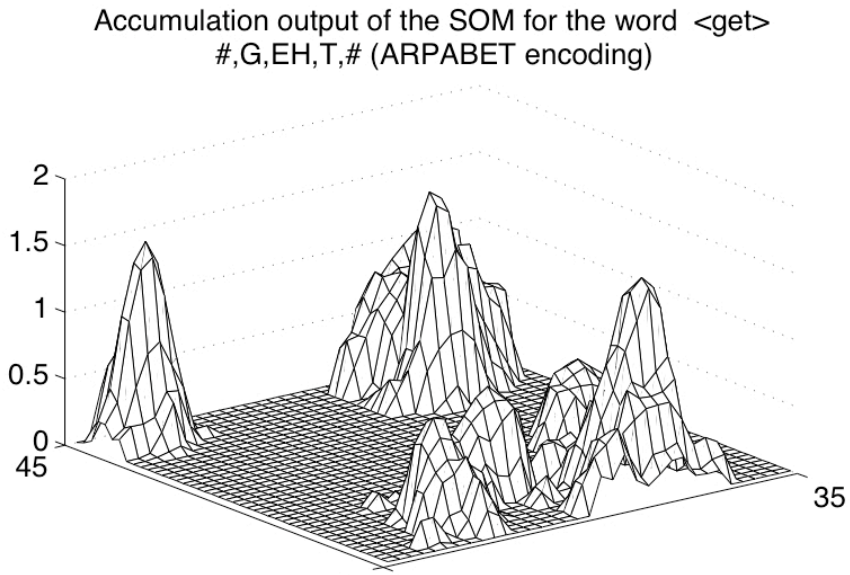


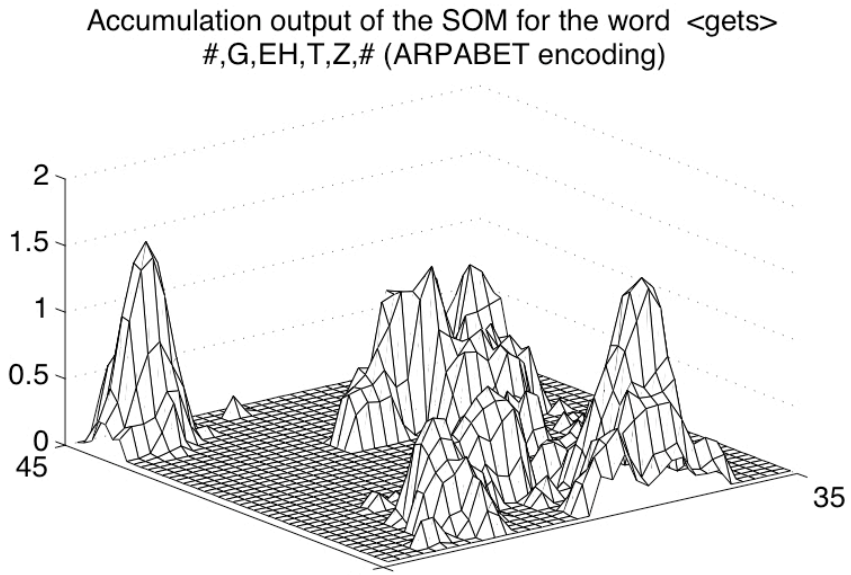
- Dans cette représentation il y a : effets de token (*n-grams* plus fréquents), effets de type (*n-grams* représentatifs de familles/series de mots), importance de la position dans le mot (#_ position initiale et _# position finale)

PHACTS : l'architecture









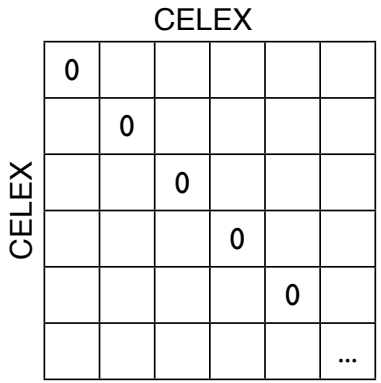
À PHACTS : distances morphologiques

Voisins pour *direction* $(1 - cos(mot_1, mot_2))$

directions *:0.0521546840 digestion :0.0583374380 vivisection :0.0664256310*
disaffection :0.0739004050 dissection :0.0805263970 discretion :0.0946790670
directional *:0.0977329500 rejection :0.1120468100 objection :0.1139618900*
defection :0.1144368500 diffraction :0.1167522700 erection :0.1199846300
affection :0.1200441600 reflection :0.1204489500 deflection :0.1210436000
election :0.1217490200 disruption :0.1222895200 recollection :0.1274858100
correction :0.1275360100 reception :0.1319253200 misconception :0.1354825300
indigestion :0.1355658600 deception :0.1356182300 differentiation :0.1376578000
dissolution :0.1386147200 electrocution :0.1390508700 subjection :0.1391268500
digression :0.1391603400 detection :0.1419080700 question :0.1428684900
misrepresentation :0.1429052900 creation :0.1457034100 exception :0.1464719800
disposition :0.1466483400 infection :0.1492515300 injection :0.1494845300
resurrection :0.1498249400 inflection :0.1503713100 prediction :0.1511868000
diminution :0.1543900400 accretion :0.1545539700 distraction :0.1565128300
precaution :0.1567157400

À PHACTS : distances morphologiques

- Pour la présente simulation :
  - Création d'une carte phonotactique à partir des *n-grams* de 30.000 mots de CELEX (pas d'information sémantique, seulement information de surface). Pas de catégories grammaticales (comme Proxinette par exemple)
  - Reconstruction des 30.000 mots comme sommation des états d'activation des *n-grams* dans la carte
  - Calcul d'une matrice de distance CELEX x CELEX (30.000 x 30.000) $\rightarrow$ métrique *cosinus*



Vue d'ensemble

Proximité morphologique

a. Deux formes qui partagent à la fois des propriétés sémantiques **et** formelles sont morphologiquement proches.

b. La proximité morphologique de deux mots est d'autant plus grande que le nombre des propriétés sémantiques et formelles qu'ils partagent est grand.

c. La proximité morphologique de deux mots est d'autant plus grande que les propriétés qu'ils partagent leur sont spécifiques, c'est-à-dire que peu d'autres mots en sont munis.

Proximité formelle

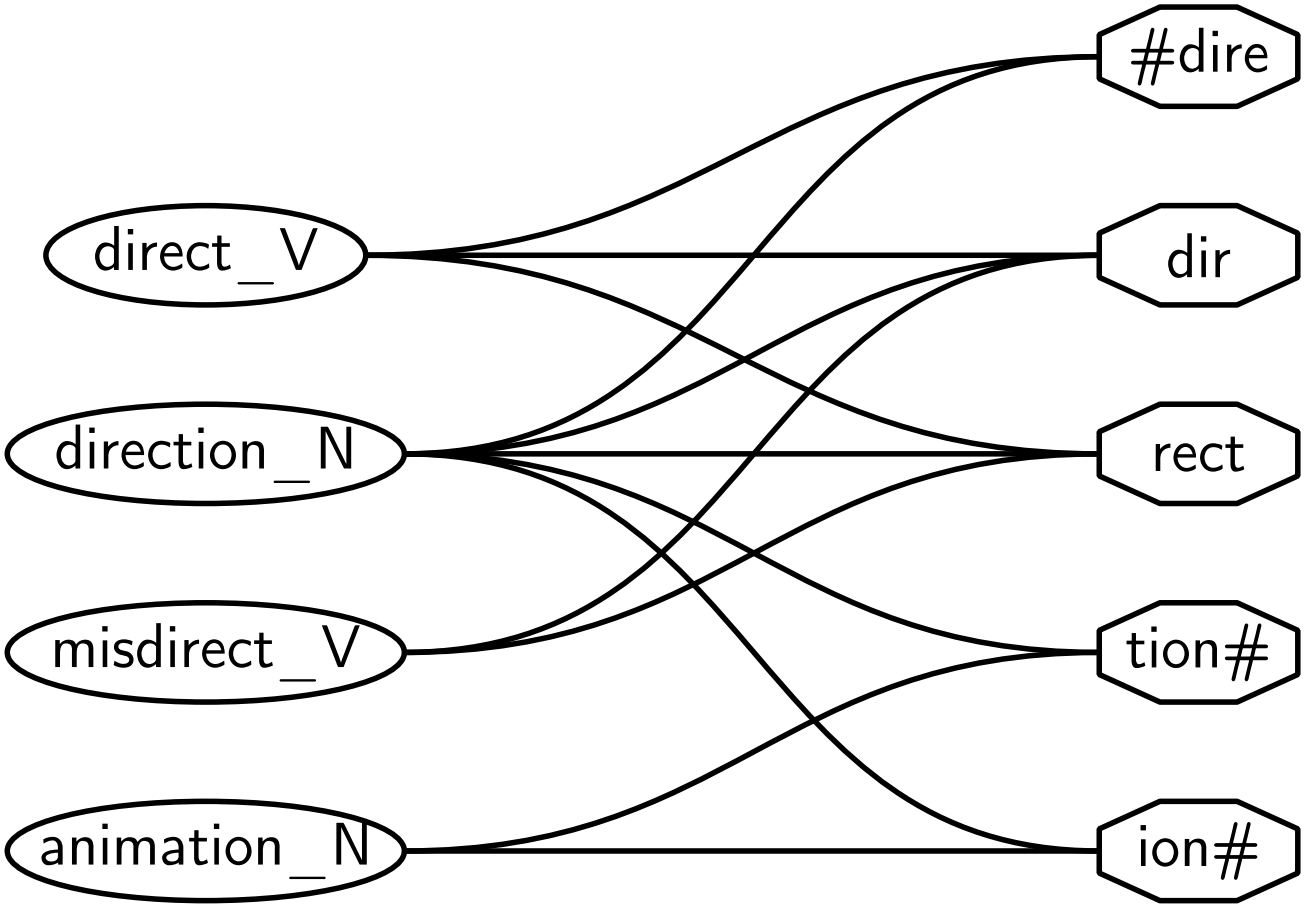
- Proxinette est une mesure de proximité formelle : elle ne prend en compte que les propriétés formelles des entrées.
- Catégorie et sens ne sont pas utilisés.
- Les propriétés sémantiques et formelles qui servent à déterminer la proximité doivent être maximalement redondantes pour capter le plus grand nombre possible de similarités qui existent entre les mots.
- Les propriétés formelles d'un mot sont l'ensemble des séquences sonores qui apparaissent dans sa forme.

Traits formels de *direction_N*

```
#direction#;
#direction; direction#;
#directio; direction; irection#;
#directi; directio; irection; rection#;
#direct; directi; irectio; rection; ection#;
#direc; direct; irecti; rectio; ection; ction#;
#dire; direc; irect; recti; ectio; ction; tion#;
#dir; dire; irec; rect; ecti; ctio; tion; ion#;
#di; dir; ire; rec; ect; cti; tio; ion; on#;
```

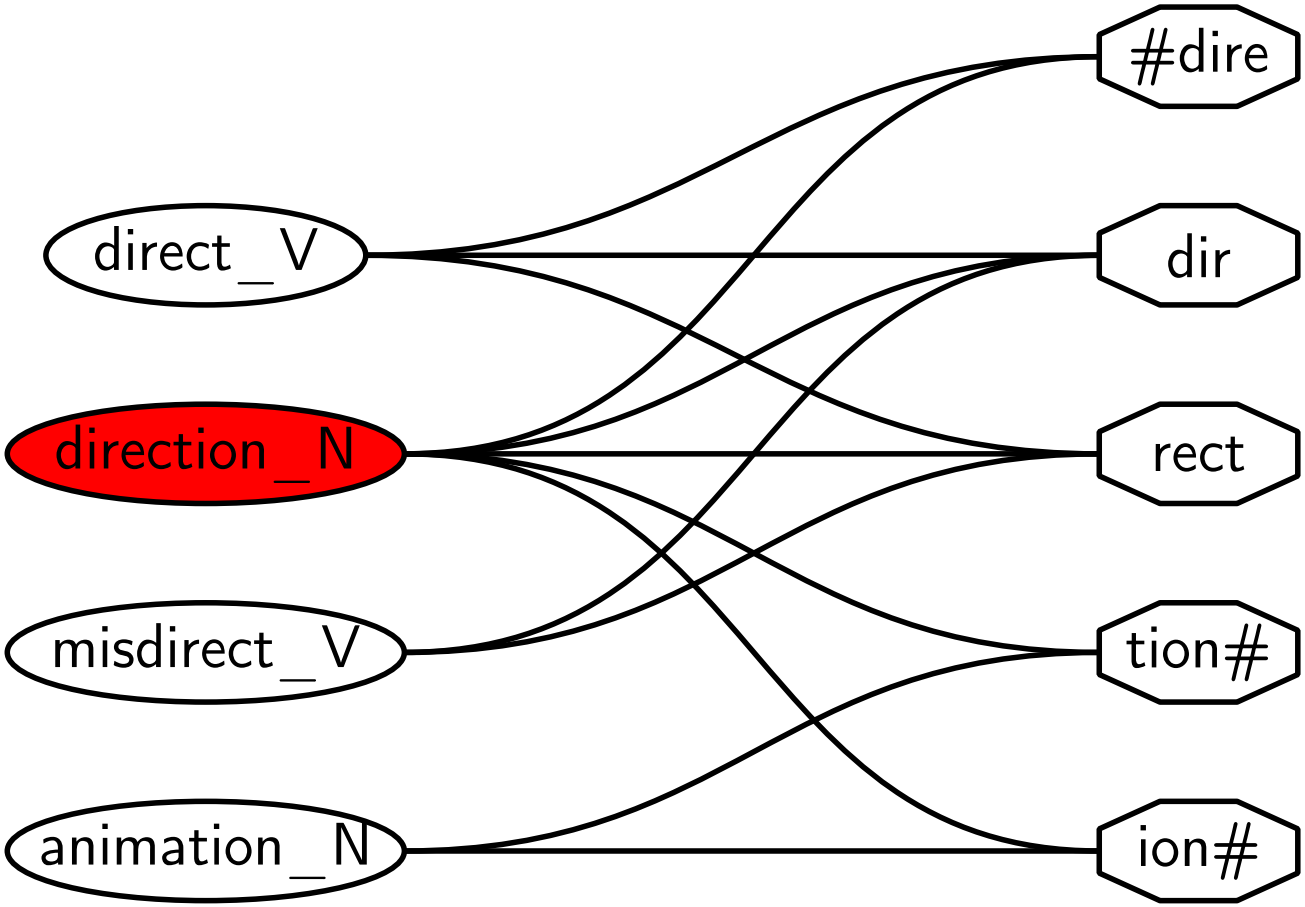
Proxinette

On construit un graphe où chaque mot est relié à l'ensemble de ses traits phonémiques.



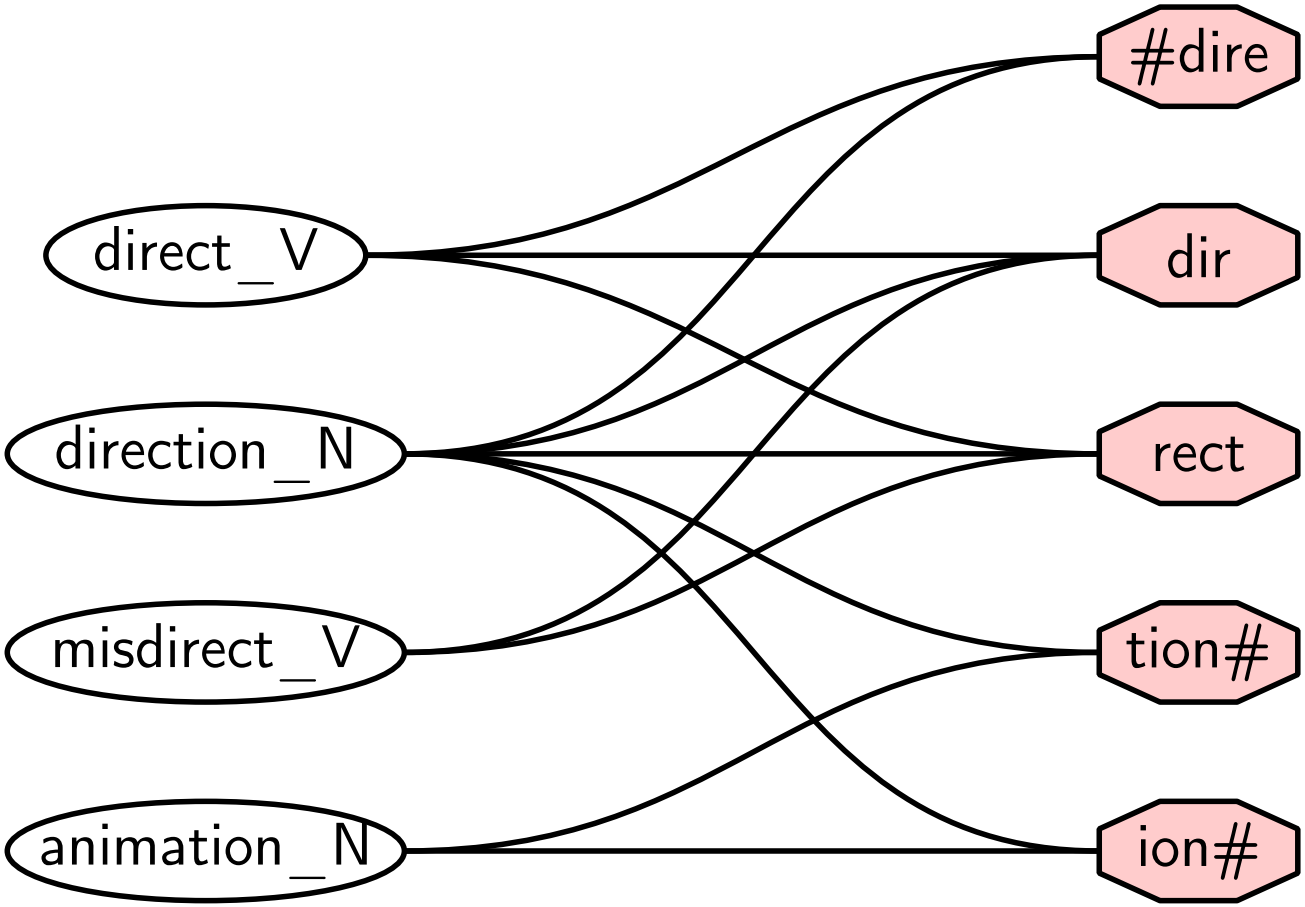
Proxinette

Une activation est générée au niveau du sommet qui représente *direction_N*.



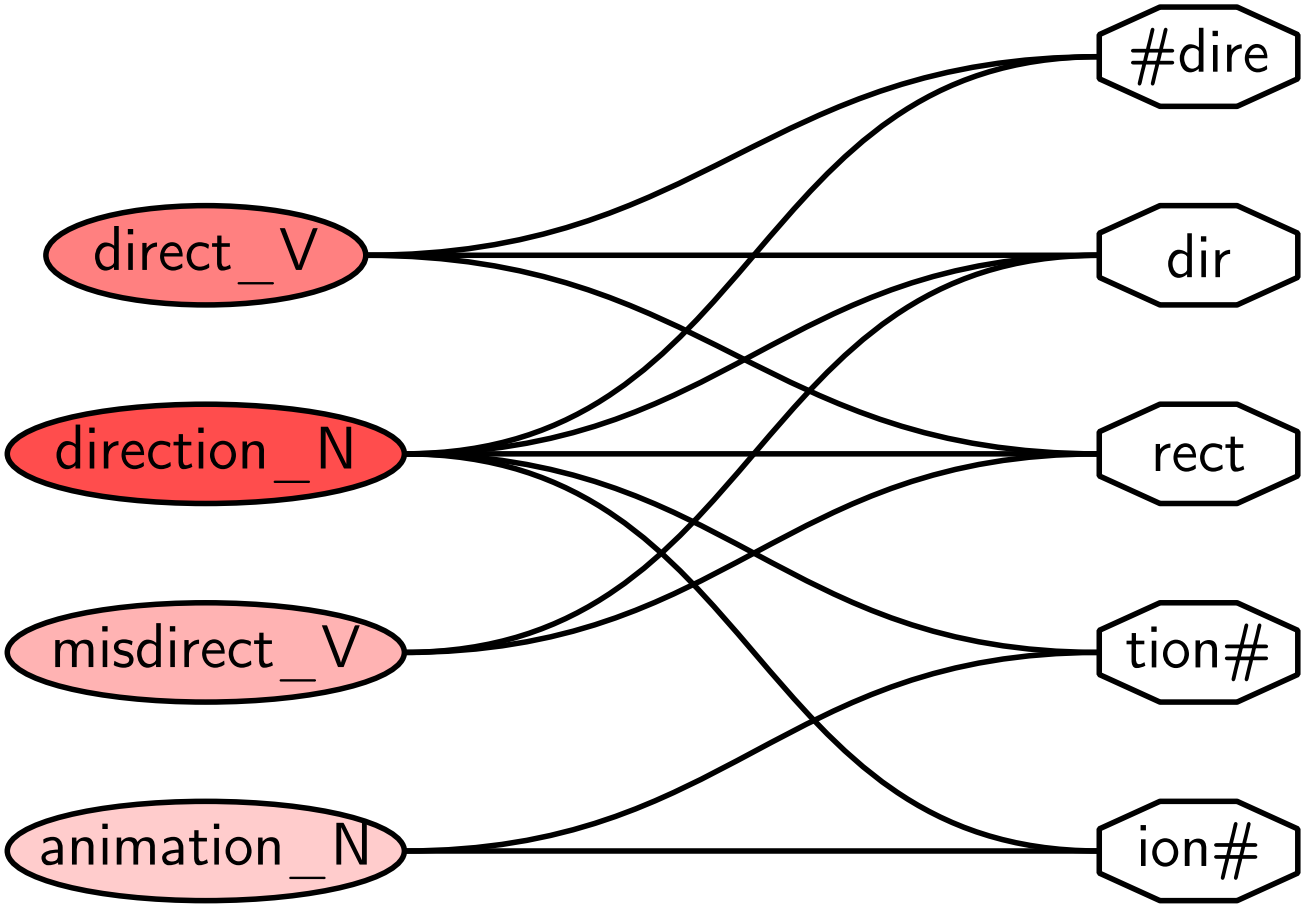
Proxinette

L'activation est propagée uniformément vers l'ensemble des traits de *direction_N*.



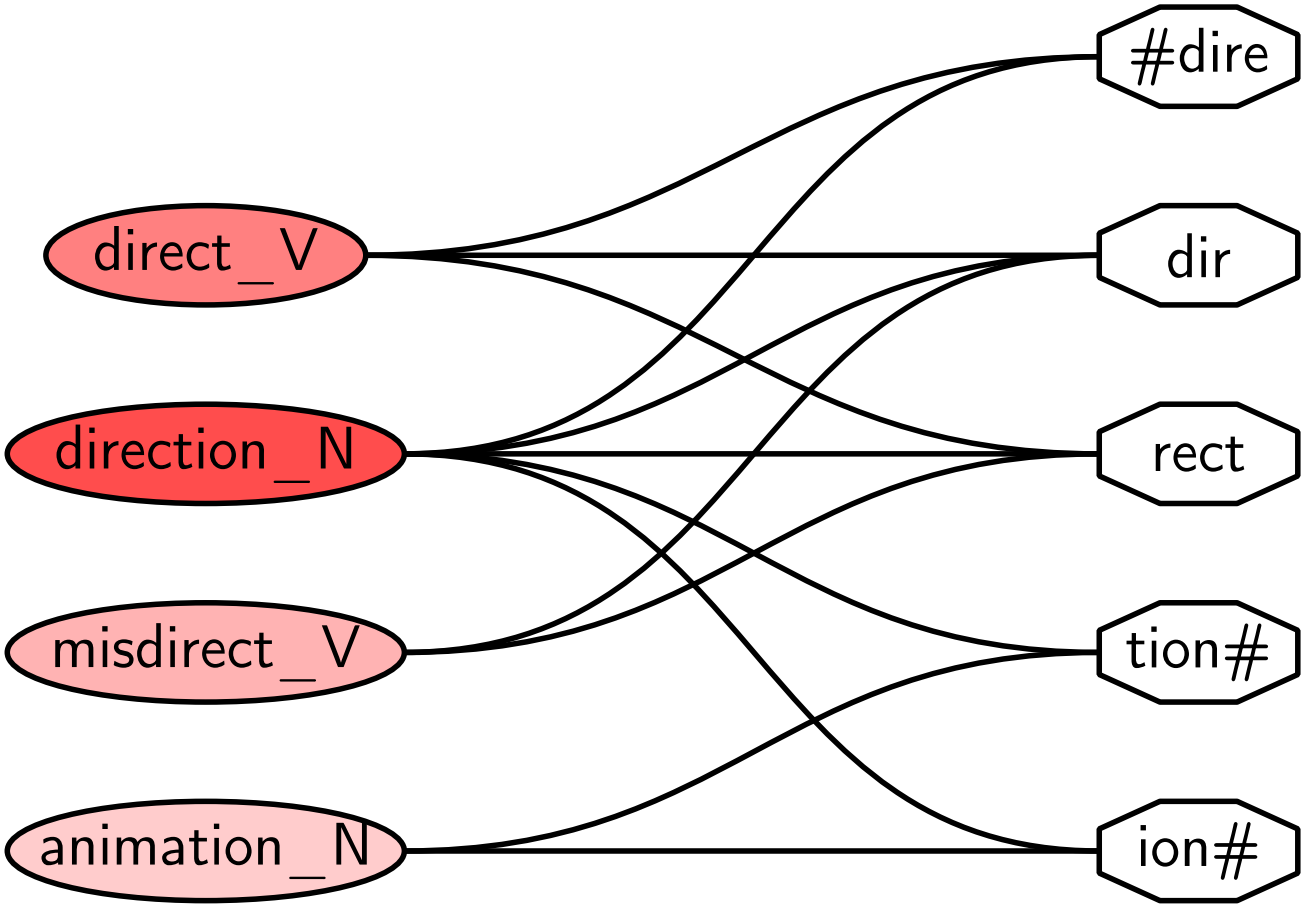
Proxinette

Les activations sont propagées uniformément vers l'ensemble des mots qui partagent des traits phonémiques avec *direction_N*.



Proxinette

Le niveau d'activation sur chaque sommet est une estimation du degré de parenté des mots correspondants avec *direction_N*.



Proxinette

Voisins de *direction_N*

direction_N :1.0000000000 directions_N :0.7359091020 directional_A :0.7359091020
misdirection_N :0.7177257909 directive_N :0.3035433276 directive_A :0.3035433276
directory_N :0.1551731447 directorship_N :0.1551731447 directorate_N :0.1551731447
director_N :0.1551731447 directness_N :0.1551731447 directly_B :0.1551731447
direct_V :0.1551731447 direct_B :0.1551731447 direct_A :0.1551731447
resurrection_N :0.1424231938 insurrection_N :0.1424231938 erection_N :0.1424231938
correction_N :0.1424231938 insurrectionist_N :0.0885551504
insurrectionism_N :0.0885551504 **redirect_V :0.0831217902 misdirect_V :0.0831217902**
indirectness_N :0.0831217902 indirectly_B :0.0831217902 indirect_A :0.0831217902
direfully_B :0.0527713005 direful_A :0.0527713005 dire_A :0.0527713005
rectification_N :0.0252537278 recollection_N :0.0251417002 dissection_N :0.0244677880
disinfection_N :0.0244677880 disconnection_N :0.0244677880 disaffection_N :0.0244677880
vivisection_N :0.0241078174 subjection_N :0.0241078174 selection_N :0.0241078174

Vue d'ensemble

Maxent

- Méthode proposée par De Pauw & Wagacha (2007)
- Mêmes traits formels que dans Proxinette : n-grammes de caractères où $3 \leq n \leq L(w)$ où $L(w)$ est la longueur du mot w
- On ajoute en plus **un trait catégoriel** (= la catégorie)

Classifieur csvLearner (Urieli)

Utilisation « créative » du classifieur :

1. On crée un modèle dans lequel il y a une catégorie différente pour chaque mot (autant de catégories qu'il y a de mots)
2. On calcule pour chaque mot la probabilité qu'il puisse appartenir à chacune des catégories du modèle.

- On n'a utilisé que les mots dont la fréquence dans le corpus COBUILD est supérieure à 10.
- Le classifieur n'utilise que les traits qui caractérisent au moins 5 classes.

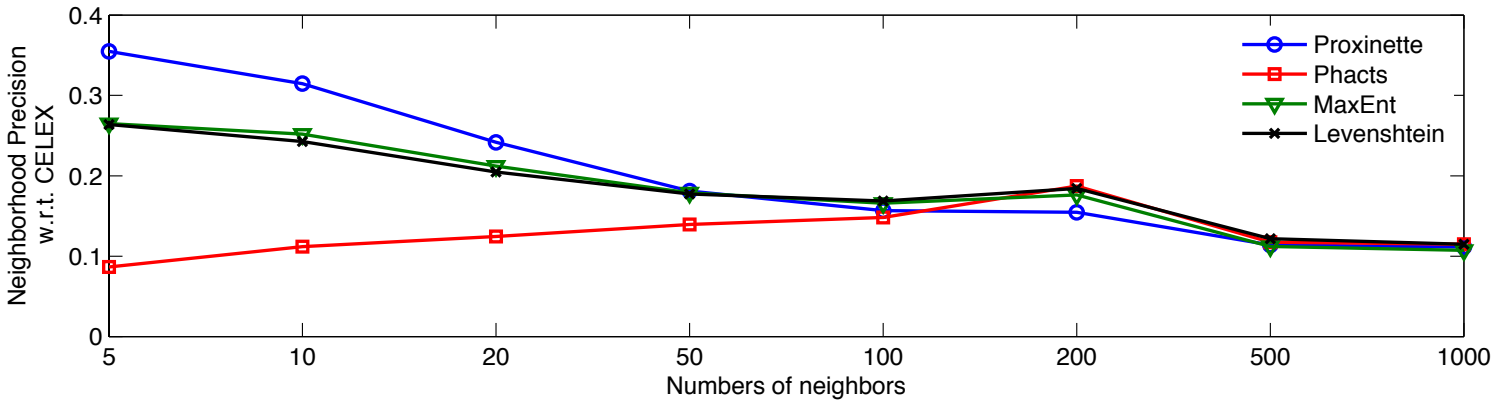
Maxent

Voisins de *direction_N*

direction_N :0.042159 directions_N :0.015270 directional_A :0.009194
directive_N :0.003306 erection_N :0.003285 resurrection_N :0.002327
correction_N :0.001961 **directness_N :0.001953 direct_V :0.001840 direct_A :0.001839**
directly_B :0.001785 director_N :0.001595 directory_N :0.001474
directorship_N :0.001119 directorate_N :0.000970 dissection_N :0.000881
vivisection_N :0.000738 inspection_N :0.000720 rejection_N :0.000716 detection_N :0.000703
defection_N :0.000701 injection_N :0.000697 connection_N :0.000683
interjection_N :0.000673 disaffection_N :0.000672 election_N :0.000669
intersection_N :0.000668 infection_N :0.000655 inflection_N :0.000644 selection_N :0.000617
deflection_N :0.000615 recollection_N :0.000586 **misdirect_V :0.000580**
redirect_V :0.000574 objection_N :0.000573 reflection_N :0.000571 affection_N :0.000570
indirect_A :0.000563 projection_N :0.000545 **indirectly_B :0.000536**
subjection_N :0.000529 protection_N :0.000529 perfection_N :0.000522
imperfection_N :0.000502 collection_N :0.000489 introspection_N :0.000443
sectional_A :0.000280 diffraction_N :0.000267 confectionery_N :0.000257

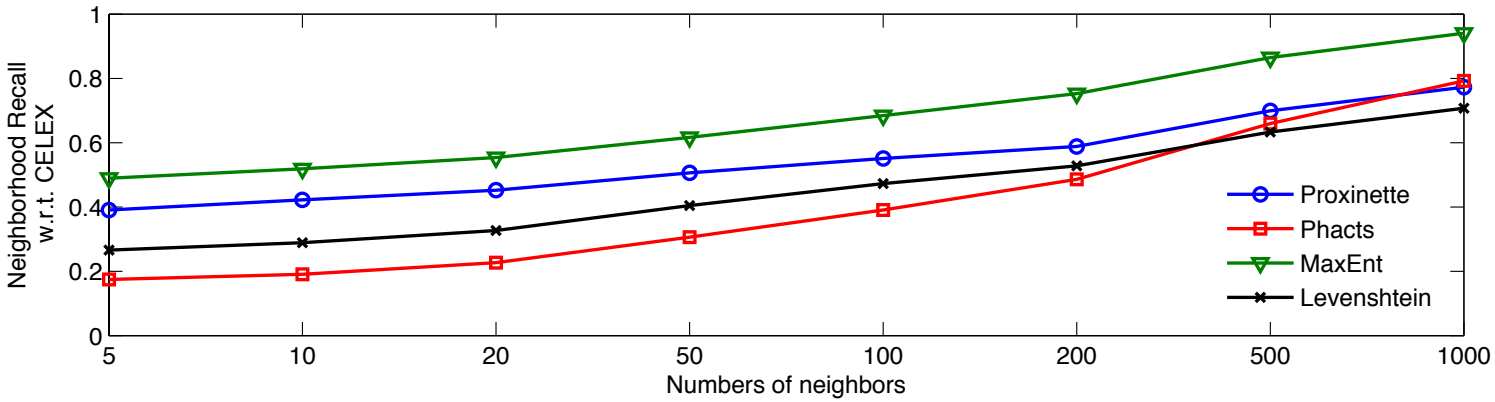
Vue d'ensemble

Precision



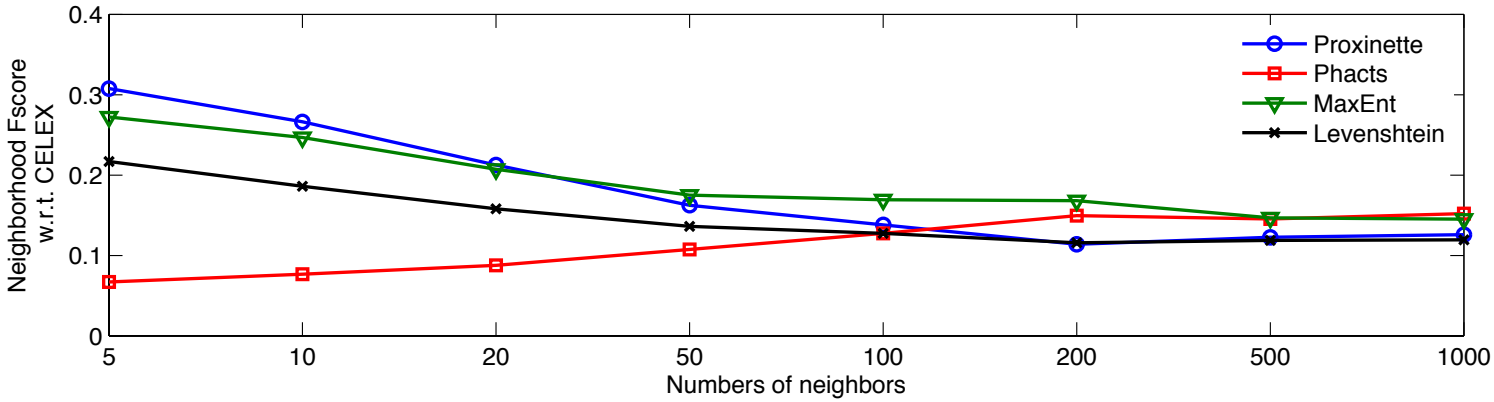
- Proxinette a le meilleur score jusqu'à 50 voisins, après il y a une convergence presque uniforme pour les 4 systèmes (0.13)
- Tendance un peu anomale pour Phacts : la seule ligne qui monte avant la convergence
  - Phacts est basé sur les *n-grams* de la langue et il développe une sensibilité statistique pour les *n-grams* très fréquents en fin de mot, comme _*ion*#, _*able*#, _*al*#, _*lly*#, (etc.) qui definissent les séries dérivationelles de chaque mot (et il sont plus nombreux des familles dérivationelles)

Recall



- MaxEnt a le meilleur rappel (0.96)
  - Ce score est très probablement lié au fait que MaxEnt est capable de capter les correspondances régulières entre catégorie et marque caractéristique des séries
  - De ce fait il est plus précis dans la sélection des éléments qui appartiennent à la série. C'est le seul qui est capable de privilégier les mots de la même catégorie

Fscore



- Phacts est le meilleur au dessus de 500 et Proxinette le meilleur en dessous de 20.
- MaxEnt s'en sort très bien et fait mieux de Proxinette au dessus de 20 parce qu'il utilise les catégories ce qui lui permet d'être plus précis dans l'identification de la série
  - Bonne performance de Phacts : il n'utilise pas des traits de catégories (comme MaxEnt)

Agenda

- Directions futures
  - Analyses plus détaillée du comportement de chaque système par rapport les analogies de CELEX : voir et isoler les approximations pour les séries et les familles
  - Pour Phacts : donner une transcription plus phonologique et pas seulement graphémique, avec un codage (binaire) des lieux/modes d'articulation
  - Pour MaxEnt : essayer une apprentissage sans les marques des catégories grammaticales, pour voir si cet aspect est directement lié à sa performance
  - Jouer avec des locuteurs : l'axe Y devrait avoir des valeurs de jugement fournies par des locuteurs sur des tâches de similarité morphologique/wordlikeness

Distribution des valeurs : Phact vs. Proxinette