Acquisition de connaissances morphologiques à partir des dictionnaires

Nabil Hathout

ERSS (CNRS & Université de Toulouse Le Mirail)

15 novembre 2005

Vue d'ensemble

Le lexique Verbaction

- Les connaissances morphologiques améliorent les performances de différentes tâches en TAL et en RI
- Le lexique Verbaction
  - 9 393 couples <verbe, nom d'action morphologiquement apparenté>

abaisser/Vmn----	abaissement/Ncms
abandonner/Vmn----	abandon/Ncms
aborder/Vmn----	abordage/Ncms
abroger/Vmn----	abrogation/Ncfs
bouder/Vmn----	bouderie/Ncfs
brocher/Vmn----	brochure/Ncfs

  - réalisé en utilisant une méthode semi-automatique :
    - génération de couples candidats en utilisant différentes ressources et différents programmes
    - révision manuelle des candidats
  - http://www.univ-tlse2.fr/erss/ressources/verbaction/main.html

Utilisations de Verbaction

- Les principales qualités de Verbaction sont
  1. qu'il ne comporte pratiquement pas d'erreurs
  2. une bonne couverture des verbes du français.
- Verbaction est utilisé par l'analyseur syntaxique Syntex pour identifier les noms d'action déverbaux et connaître le verbe qui leur correspond.
  - Ces informations servent à améliorer le rattachement prépositionnel à ces noms.
  - Elles permettent d'utiliser les informations de sous-catégorisations apprises pour les verbes comme indices pour désambiguiser ces rattachements.
- Verbaction a aussi été utilisé en recherche d'information pour l'expansion de requêtes. Expériences réalisées sur la plateforme RFIEC (Ludovic Tanguy et Marianne Vergez).

Ressources morphogiques « sures »

- Les ressources morphologiques comme Verbaction sont rares car elles sont difficiles à construire
- Il n'existe pas pour le français une base de données morphologique similaire à la base Celex.
  Une version en ligne de Celex est consultable sur
  http://www.mpi.nl/world/celex/
- La seule ressource morphologique à large couverture existante est le dictionnaire DISFA constitué par Claude Gruaz et son équipe et qui est diffusé en format PDF.
  http://www.up.univ-mrs.fr/delic/disfa/accueil.html

Traitements morphologiques par règles

- Les connaissances morphologiques sont incluses dans les programmes
- Les analyseur morphologiques comme l'analyseur DériF développé par Fiammetta Namer.
  - L'analyse d'un lexème permet de déterminer s'il est simple ou dérivé et dans le second cas, de le décomposer en un affixe et un radical
- Le raciniseur de Porter (largement utilisé en RI pour des langues à morphologie flexionnelle pauvre comme l'anglais)
- La troncation (à 6 ou 7 caractéres)

Méthodes par apprentissage

Ces méthodes sont bien adaptées à la construction semi-automatique de ressources morphologiques.

1. Observer des sources « sures » comme des lexiques flexionnels ou des corpus pour découvrir des régularités morphologiques
2. Utiliser ces régularités pour construire des ressources morphologiques (lexiques morphologiques ; bases de données morphologiques)
3. Ces ressources plus ou moins bruitées peuvent/doivent ensuite être révisées
    - Même si les ressources résultats ne contiennent que très peu d'erreurs, ces dernières sont très pénalisantes pour les applications qui les utilisent : la dégradation des résultats due à ces erreurs dépassent généralement les gains apportés par la partie correcte de ces ressources.

Méthodes par apprentissage

- Les outils permettant de constituer des ressources de ce genre sont nombreux (Projet « Linguistica » de Goldsmith basé sur les MLD ; travaux de Schone et Jurafsky basés sur les LSA...).
- Tous ces outils utilisent des méthodes statistiques plus ou moins sophistiquées pour identifier les régularités « pertinentes » : MLD, LSA, variantes de l'information mutuelle...
- Les méthodes par apprentissage permettent de constituer des ressources à large couverture : elles ne se limitent pas aux lexèmes dont une analyse est implémentée dans un analyseur.
- Elles sont largement indépendantes vis-à-vis des langues particulières

Objectifs généraux

- Constituer **semi-automatiquement** des bases de données dérivationnelles pour le français
  - destinées au TAL, à la linguistique de corpus, à la recherche d'information, à la psycholinguistique...
  - un processus actif
    l'activité du processus
    activer un processus
    un processus activable
    activation des processus
- Utilisation de ressources lexicographiques « externes » mais pas de connaissances linguistiques dans les outils.
  1. Lexiques flexionnels (Morphalou, ABU, TLFnome)
  2. Dictionnaires de synonymes (Dicosyn, WordNet)
  3. Dictionnaires de langues (Robert, TLFi)

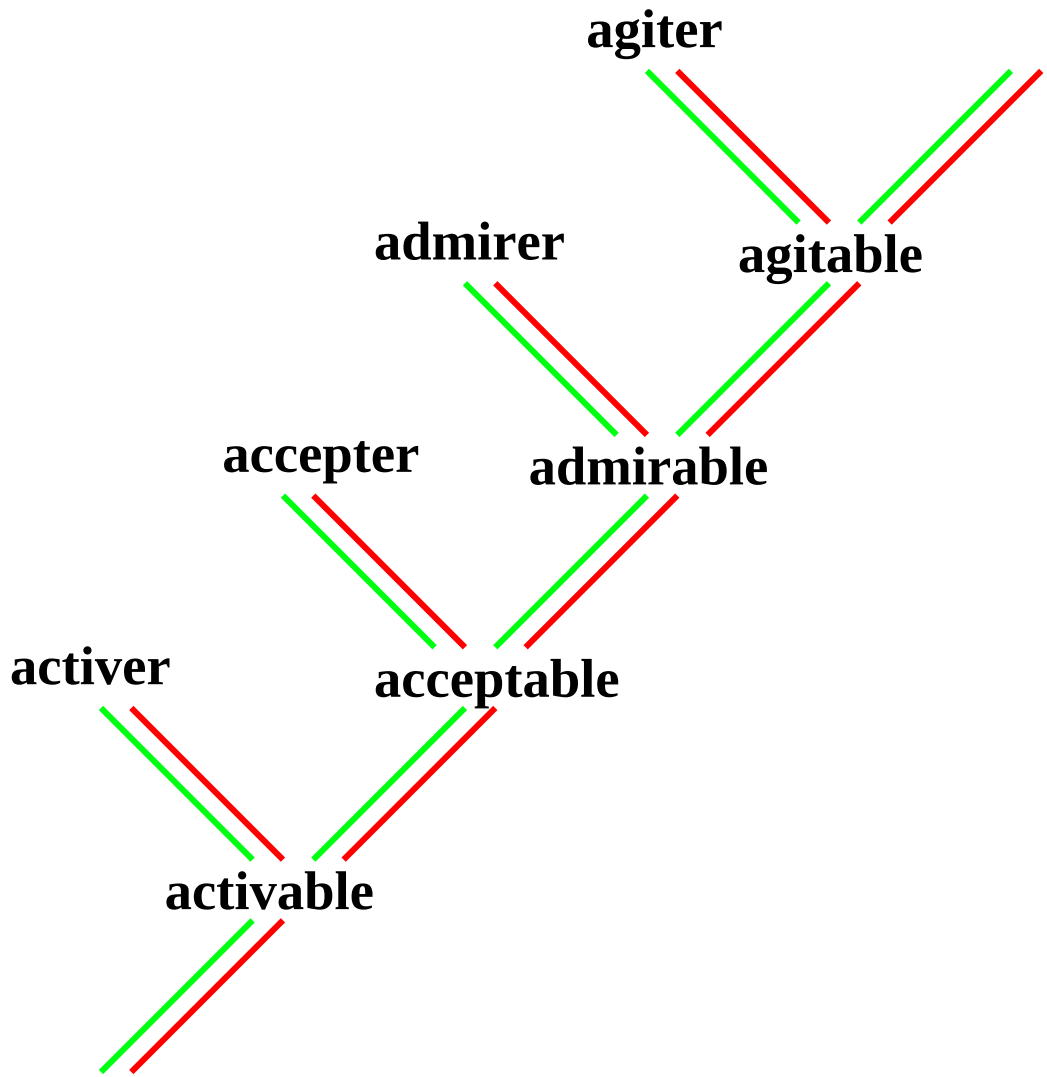
La structure du lexique

- Le **lexique** est un **graphe**. C'est un ensemble de formes (fléchies) attestées connectées en fonction des propriétés sémantiques et phonologiques qu'elles partagent (Bybee 1985, 1995, 1998) ;
- La **dérivation morphologique** est une **relation** entre des lexèmes qui partagent à la fois des propriétés sémantiques et des propriétés phonologiques ;
- Un **affixe dérivationnel** (préfixe ou suffixe) est un sous-graphe qui forme une **série proportionnelle** et que l'on peut étendre par analogie ;
- Un **lexème** est un ensemble de formes de même catégorie qui ne diffèrent que leurs marques flexionnelles.

Morphologie dérivationnelle

partage de sens

partage de sons



Morphologie dérivationnelle

- La dérivation des lexèmes s'appuie sur :
  - les séries (organisation du lexique existant)
  - l'analogie (pour ajouter de nouveaux couples aux séries existantes)
- Les relations morphologiques ne s'établissent pas seulement entre les bases et leurs dérivés. Les relations de partage de son et de sens existent aussi entre
  - les lexèmes construits au moyen d'un même affixe
  - les lexèmes qui appartiennnent à une même famille dérivaitionnelle... *auditeur* et *audition* sont morphologiquement apparentés même s'ils n'ont pas de base attestée en français (pas de verbe °*audire*)

Analogie graphémique

- Graphémiquement :

  activable/Afpms est à activer/Vmn----
  ce que agitable/Afmps est à agiter/Vmn----

  Dans les 2 coupes able/Afpms est remplacé par er/Vmn----.

  - étiquettes morphosyntaxiques Grace/Multext/Eagle ;
  - Afpms = adjectif qualificatif positif masculin singulier ;
  - Vmn---- = verbe non auxiliaire infinitif.

- L'analogie définit une structure sur le lexique

activable/Afpms	activer/Vmn----
agitable/Afmps	agiter/Vmn----
accordable/Afpms	accorder/Vmn----
achetable/Afpms	acheter/Vmn----

  Elle est à l'origine de l'agrégation des relations morphologiques en séries.

Schémas de suffixation

- On veut constituer des séries morphographiques à partir de lexiques flexionnels, c'est-à-dire de listes de mots munis d'étiquettes morphosyntaxiques (ex. Morphalou, ABU, Multext, TLFnome...)
- Hypothèse : les lexèmes suffixés sont de la forme :

  radical · suffixe

  Par exemple : activ · able ; accord · er...
- On associe à chaque couple de formes $(m_1, m_2) \in L^* \times L^*$ un ensemble $S(m_1, m_2)$ de couples de suffixes, défini comme suit :

$$S(m_1, m_2) = \{(t_1, t_2) \in L^* \times L^* / \exists r \in L^*, m_1 = r \cdot t_1 \text{ et } m_2 = r \cdot t_2\}$$

- Exemple : $S(\texttt{laver}, \texttt{lavable}) = \{(\texttt{laver}, \texttt{lavable}), (\texttt{aver}, \texttt{avable}), (\texttt{ver}, \texttt{vable}), (\texttt{er}, \texttt{able})\}$
- $S(\texttt{fixation}, \texttt{fixable}) = \{(\texttt{fixation}, \texttt{fixable}), (\texttt{ixation}, \texttt{ixable}), (\texttt{xation}, \texttt{xable}), (\texttt{ation}, \texttt{able}), (\texttt{tion}, \texttt{ble})\}$

Signature suffixale

- La signature suffixale $\sigma(m_1, m_2)$, est l'élément $(s_1, s_2) \in S(m_1, m_2)$ pour lequel r est de longueur maximale.
- $\sigma(\texttt{laver}, \texttt{lavable}) = (\texttt{er}, \texttt{able})$
- $\sigma(\texttt{fixation}, \texttt{fixable}) = (\texttt{tion}, \texttt{ble})$
- En ajoutant les étiquettes morphosyntaxiques : able/Afpms:er/Vmn---- ; tion/Ncfs:ble/Afpms
- La signature suffixale caractérise le rapport graphémique entre les membres du couple de lexèmes. La signature d'un couple est appelée son « schéma de suffixation »
- On peut ainsi abstraire un schéma de suffixation pour chaque couple de formes

Utilisation des schémas de suffixation

- Les relations morphographiques représentées par les schémas de suffixation permettent de connecter certains lexèmes du lexique.
- Par exemple able/Afpms:er/Vmn---- permet de connecter deux lexèmes X et Y si :
  - X porte l'étiquette Afpms ;
  - Y porte l'étiquette Vmn---- ;
  - Y peut être formée à partir de X en lui retirant le suffixe graphémique -able et en lui ajoutant le suffixe graphémique -er.
- L'ensemble des couples qui ont la même signature forment une **série morphographique**.
- Comment identifier les séries morphographiques qui sont pertinentes linguistiquement ?

Apprentissage de schémas de suffixation

- Le partage d'un radical graphémique par X et Y est
  - une **approximation** du partage de propriétés phonologiques **et**
  - une **approximation** du partage de propriétés sémantiques
- Paramètres de l'apprentissage :
  - La taille minimale du radical doit dépasser un seuil déterminé (en général 3)
  - La taille maximale de l'affixe ne doit pas dépasser un seuil déterminé (en général 8) pour ne pas conserver des schémas qui relient des formes préfixées ou partageant un même élément de composition (*métrique*, *morphique*, *dynamique*...)
  - La **fréquence lexicale** du schéma de suffixation est la taille de sa série morphographique. C'est le nombre de couples du lexique d'apprentissage qui peuvent être connectés par le schéma
    - La **fréquence lexicale** est un indice de régularité du schéma et un gage de sa validité
    - On impose en général une fréquence lexicale minimale entre 5 et 8

DéCor

- DéCor : outil permettant d'apparier des lexèmes dérivés avec leurs lexèmes bases

1. extraction de 2 ensembles de lexèmes : l'ensemble des dérivés et un ensemble de bases potentielles
2. apprentissage de schémas de suffixation
3. constitution de séries morphographiques en appliquant les schémas de suffixation : constitution de couples candidats
4. filtrage des candidats par la fréquence lexicale des schémas : on ne garde pour chaque lexème dérivé que le couple qui appartient à la série la plus grosse.

DéCor

- Précision calculée relativement aux analyses de DériF qui ont été révisées par Georgette Dal et Fiammetta Namer.
- La précision des résultats dépend de l'homogénïté catégorielle des bases.

suffixe	*-able*	*-ité*	*-iser*
précision	94%	89%	20%

- Les bases des adjectifs en *-able* sont presque toutes des verbes (très peu de verbes).
- Les bases des noms en *-ité* sont en grande majorité des adjectifs (une petite proportion de noms comme *auteur*>*autorité*).
- Les bases des verbes en *-iser* se composent de noms et d'adjectifs en proportion équilibrée (*atome*>*atomiser* ; *axiomatique*>*axiomatiser*).

DéCor et DéClique

- DéCor souffre de deux handicaps :
  1. Les schémas de suffixation connectent des lexèmes morphologiquement apparentés et pas seulement les lexèmes dérivés à leurs lexèmes bases
  2. DéCor n'utilise que les graphies or **la dérivation morphologique est d'abord une affaire de sémantique**
- **DéClique** : classification en familles dérivationnelles (en cliques)
  - fraternaliste/Afpms ; fraternaliste/Ncms ; fraternisation/Ncfs ; fraternellement/Rgp ; fraternel/Afpms ; fraterniser/Vmn---- ; fraternité/Ncfs
  - freinateur/Afpms ; freinateur/Ncms ; freination/Ncfs ; freinage/Ncms ; freinette/Ncfs ; freineur/Ncms ; freiner/Vmn---- ; freinée/Ncfs ; frein/Ncms
  - Évaluation sur la tranche *fr-* du TLF : précision : 80% ; rappel : 60%

Dictionnaires de synonymes

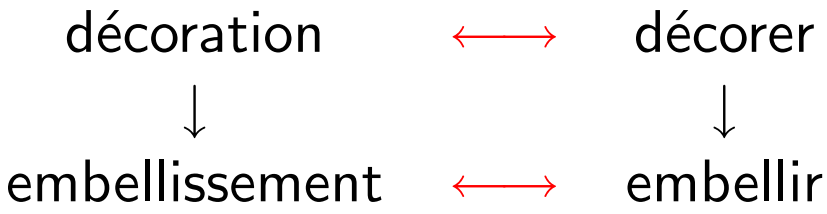
- L'approximation par la graphie :
  - ✔ bonne pour les propriétés phonologiques
  - ✘ insuffisante pour les propriétés sémantiques
- Les dictionnaires de synonymes fournissent précisément les informations sémantiques dont la morphologie a besoin :
  - La synonymie (ou la proximité sémantique) est une relation de partage de propriétés sémantiques
  - Mais pas de relation entre des lexèmes morphologiquement apparentés
- On doit utiliser les dictionnaires de synonymes de manière indirecte pour filtrer les couples morphographémiques

Quadruplets morpho-synonymique

- On forme des quadruplets morpho-synonymiques $X_1{:}X_2{::}Y_1{:}Y_2$ tels que :
  1. $X_1{:}X_2$ et $Y_1{:}Y_2$ sont morphologiquement apparentés (relations de suffixation graphémique prédites)
  2. X_1 est un synonyme de Y_1 (relation donnée par le dictionnaire)
  3. X_2 est un synonyme de Y_2 (relation donnée par le dictionnaire)
- Si les relations morphologiques prédites sont correctes alors $X_1{:}X_2{::}Y_1{:}Y_2$ est un quadruplet analogique :

« X_1 est à X_2 ce que Y_1 est à Y_2 »

Exemple de quadruplet



- Ce quadruplet permet d'acquérir 2 couples :
  - *décoration*:*décorer*
  - *embellissement*:*embellir*

Quelques filtres supplémentaires

1. Les formes graphémiques de X_1, X_2, Y_1 et Y_2 doivent toutes être différentes
   anis/Ncms:*anisette/Ncfs*::*anisade/Ncfs*:*anis/Ncms*
2. Seules les relations morphographiques et/ou synonymiques contrôlées par les contraintes de la définition peuvent s'établir entre les éléments du quadruplet
   forger/Vmn----:*former/Vmn----*::
   constituer/Vmn----:*construire/Vmn----*
3. Le radical graphémique de X_1:X_2 ne doit être ni un préfixe ni un suffixe dans celui de Y_1:Y_2 et réciproquement
   bassin/Ncms:*bassinage/Ncms*:: *bassinoire/Ncfs*:*bassinement/Ncms*

Expérience sur le français

- Nous avons utilisé le dictionnaire de synonymes DICOSYN constitué à l'INaLF/ATILF :
  compilation de 5 dictionnaires de synonymes (R. Bailly, H. Bénac, H. Bertaud du Chazaud, F. Guizot, B. Lafaye) + synonyme du *Grand Larousse* + renvois analogiques du *Grand Robert*
- Formatage en une liste de couples de lexèmes :
  - suppression des locutions et des syntagmes
  - étiquetage morphosyntaxique au moyen de TLFnome
    (suppression les synonymes qui ne sont pas de la même catégorie que l'entrée)
  - pas de symétrisation des relations de synonymies
- 45 009 lexèmes (entrées ou synonymes)
- 234 771 relations de synonymie (proximité sémantique)

Exemples de quadruplets acquis

✔ alchimie/Ncfs : alchimiste/Ncms :: archimagie/Ncfs : archimage/Ncms
✔ alternance/Ncfs : alternant/Afpms :: récurrence/Ncfs : récurrent/Afpms
✘ facétie/Ncfs : facétieux/Afpms :: drôlerie/Ncfs : drôle/Afpms
✔ fouiller/Vmn---- : fouilleur/Afpms :: fureter/Vmn---- : fureteur/Afpms
✔ introniser/Vmn---- : intronisation/Ncfs ::
couronner/Vmn---- : couronnement/Ncms
✔ métaphore/Ncfs : métaphorique/Afpms ::
symbole/Ncms : symbolique/Afpms
✘ rigoler/Vmn---- : rigolade/Ncfs :: blaguer/Vmn---- : blague/Ncfs
✔ révéler/Vmn---- : révélateur/Ncms ::
divulguer/Vmn---- : divulgateur/Ncms
✔ sobriété/Ncfs : sobre/Afpms :: simplicité/Ncfs : simple/Afpms
✔ toucher/Vmn---- : touchant/Afpms ::
troubler/Vmn---- : troublant/Afpms

Évaluation des résultats

filtres suppl.	quad.	couples	préc. (%)	rappel (%) Verbaction	*-fr*
aucun	47 426	26 870	85,5	93,4	63,4
1	43 377	24 499	**95,0**	93,3	65,1
2	39 474	20 861	89,0	97,4	65,5
3	40 535	21 929	94,0	96,6	61,0
1et 3	36 649	19 581	92,0	96,5	62,0
1, 2 et 3	35 208	18 286	89,5	97,4	69,0

Les contraintes supplémentaires (2) et (3) dégradent les performances car elles éliminent plus de quadruplets corrects que de quadruplets incorrects.

Expérience sur l'anglais

Objectifs de l'expérience :

1. Vérifier l'**indépendance** de la méthode vis-à-vis des langues particulières
2. Vérifier la **robustesse** de la méthode (en utilisant des relations de proximité sémantique de plus en plus lâches)
3. Montrer que la méthode est **utile** pour les langues qui disposent déjà de bases de données morphologiques (CELEX : anglais ; néerlandais ; allemand) (Baayen et al. 1995)

WordNet

The verb copy has 4 senses (first 2 from tagged texts)

1. (3) **copy**#1 – (copy down as is; "The students were made to copy the alphabet over and over")
2. (3) imitate#1, **copy**#2, simulate#1 – (reproduce someone's behavior or looks; "The mime imitated the passers-by"; "Children often copy their parents or older siblings")
3. imitate#2, **copy**#3 – (imitate in behavior or appearance; "She is imitating the comedian very well!")
4. **copy**#4, re-create#2 – (make a replica of; "copy that drawing"; "re-create a picture by Rembrandt")

Dictionnaires de synonymes anglais

3 dictionnaires de relation de proximité sémantique ont été extraits de WordNet 1.7 :

1. S-dict : appartenance à un même synset (proximité forte)
2. M-dict : S-dict + appartenance à des synsets synonymes et hyperonymes immédiats (proximité moyenne)
3. L-dict : M-dict + apparetenance à des synsets co-hyponymes (proximité faible)

		# lemmes
S-dict	127 274	43 055
M-dict	283 422	62 477
L-dict	2 213 331	64 168

Expérience sur l'anglais

- Acquisition réalisée avec exactement les mêmes programmes que pour l'expérience sur le français

	acquisition		évaluation	
	# quad.	# couples	précis.	rappel
S-dict	41 117	24 079	**92 %**	78,6 %
M-dict	77 132	43 349	86 %	77,5 %
L-dict	334 516	73 376	62 %	72,3 %

Typer les quadruplets

On peut distinguer 2 types de quadruplets :

1. Quadruplets dont la signature est exacte (E) :
   la signature de X_1:X_2 est identique à celle de Y_1:Y_2
   adorer/Vmn---- : adoration/Ncfs ::
   vénérer/Vmn---- : vénération/Ncfs
2. Quadruplets dont la signature est hétérogène (H) :
   la signature de X_1:X_2 est différente de celle de Y_1:Y_2
   introniser/Vmn---- : intronisation/Ncfs ::
   couronner/Vmn---- : couronnement/Ncms

Typage des schémas d'affixation

Le typage des quadruplets induit un typage des schémas :

1. Un schéma est exact (E) si il apparaît dans **au moins un** quadruplet exact
   er/Vmn----:ation/Ncfs (adorer:adoration)
2. Un schéma est hétérogène (H) s'il n'apparaît **que** dans des quadruplets hétérogènes
   r/Vmn----:on/Ncfs (réunir:réunion)
   /Ncms:leux/Afpms (péril:périlleux)

Sous-types de quadruplets

Affiner le typage des quadruplets

1. E
   haut/Ncms:hauteur/Ncfs :: profond/Ncms:profondeur/Ncfs
2. H0 = E:E
   onduleux/Afpms:ondulement/Ncms ::
   ondulant/Afpms:ondulation/Ncfs
3. H1 = E::H ou H::E
   égaliser/Vmn----:égal/Ncms :: unifier/Vmn----:uni/Ncms
4. H2 = H::H
   suspension/Ncfs:suspendre/Vmn---- :: fermeture/Ncfs:fermer/Vmn----

Types de quadruplets

- Statistiques pour DICOSYN :

type	#signature	#quadruplets	moyenne
E	567	8 160	14,4
H	18 951	35 217	1,9
H0	2 231	9 183	4,1
H1	9 117	17 222	1,9
H2	7 603	8 812	1,2

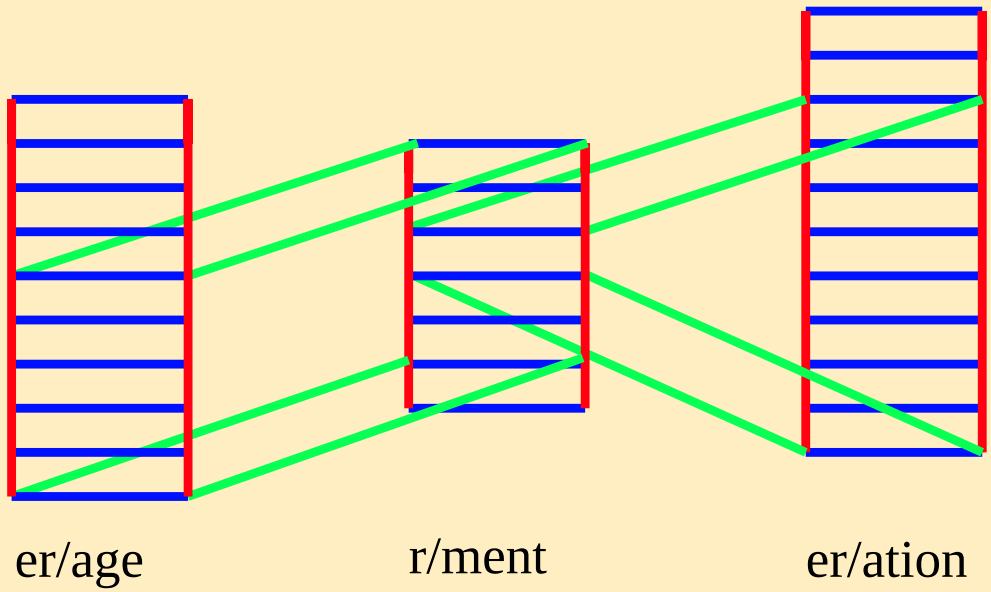
- Statistiques pour le dictionnaire *S-dict* :

type	#signature	#quadruplets	moyenne
E	825	8 014	9,7
H	16 000	33 103	2,1
H0	1 640	4 924	3,0
H1	6 962	16 838	2,4
H2	7 398	11 341	1,5

Analogie et structuration du lexique

- Les quadruplets exacts forment les séries proportionnelles les plus nombreuses
- Les quadruplets H0 forment des petites séries
- Les quadruplets H1 et H2 forment des séries « minimales »

couples morpho.
quad. E
quad. H0



Sous-types de schémas des qudruplets H1

Affiner encore le typage des schémas :

- H1-E le schéma E d'un quadruplet H1
  égaliser/Vmn----:égal/Ncms :: unifier/Vmn----:uni/Ncms
- H1-H le schéma H d'un quadruplet H1
  égaliser/Vmn----:égal/Ncms :: unifier/Vmn----:uni/Ncms

Précision et rappel par types d'analogie

- DICOSYN :

type	#cpl	précision	rappel V	rappel *fr-*
E	7 385	**98,0 %**	97,7 %	73,3 %
H0	6 652	**98,5 %**	96,5 %	76,3 %
H1-E	10 822	**99,0 %**	94,9 %	59,7 %
H1-H	14 876	78,5 %	84,5 %	59,1 %
H2	9 341	81,0 %	76,8 %	52,9 %

- La diversité des contraintes est plus déterminante que leur force (le nombre de couples qui forment la série) :
  - Les quadruplets E ne sont contraints que dans **1** série
  - Les quadruplets H0 sont contraints dans **2** séries
  - Les quadruplets H1-E sont contraints dans **2** séries, dont l'une est elle-même contraintes dans **2** séries

Précision et rappel par types d'analogie

S-dict :

type	#cpl	précision	rappel
E	6 932	94,5 %	92,1 %
H0	3 734	**95,5 %**	87,8 %
H1-E	9 398	**99,5 %**	80,2 %
H1-H	10 539	**96,0 %**	77,5 %
H2	9 123	85,0 %	65,9 %

Précision et rappel par types d'analogie

M-dict :

type	#cpl	précision	rappel
E	12 280	93,0 %	89,2 %
H0	10 500	**97,5 %**	83,7 %
H1-E	16 383	**97,0 %**	79,2 %
H1-H	17 868	91,0 %	72,3 %
H2	16 622	72,5 %	55,2 %

L-dict :

type	#cpl	précision	rappel
E	18 006	81,0 %	84,3 %
H0	18 635	90,5 %	78,4 %
H1-E	27 177	82,5 %	73,7 %
H1-H	35 275	68,5 %	62,9 %
H2	40 933	46,0 %	49,7 %

Acquisition à partir de dictionnaires de langue

- Calculer les partages de sens en utilisant un algorithme comme PROX qui permet de mesurer la distance sémantique entre des lexèmes
- Croiser le graphe morphographique et le graphe des proximités sémantiques
  - Le croisement sera direct ; inutile de passer par l'analogie
  - La principale difficulté : le graphe morphographique est booléen ; celui des proximités sémantiques est valué sur les réels

Acquisition à partir de dictionnaires de langue

- Est-il possible « d'affaiblir » les relations morphographiques pour prendre en compte les allomorphies, les conversions... ?
  Comment donner la primauté à la sémantique ?
- Identifier des séries morphologiques à partir de sous-structures dans le graphe des proximités sémantiques.
  - Quels sont les sous-graphes qui caractérisent les relations entre les noms d'action en *-ion* et leurs bases ?
  - Application linguistique directe : la comparaison d'affixes concurrents *-age*, *-ment*, *-tion*...