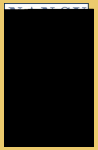
Unité mixte de recherche 7118



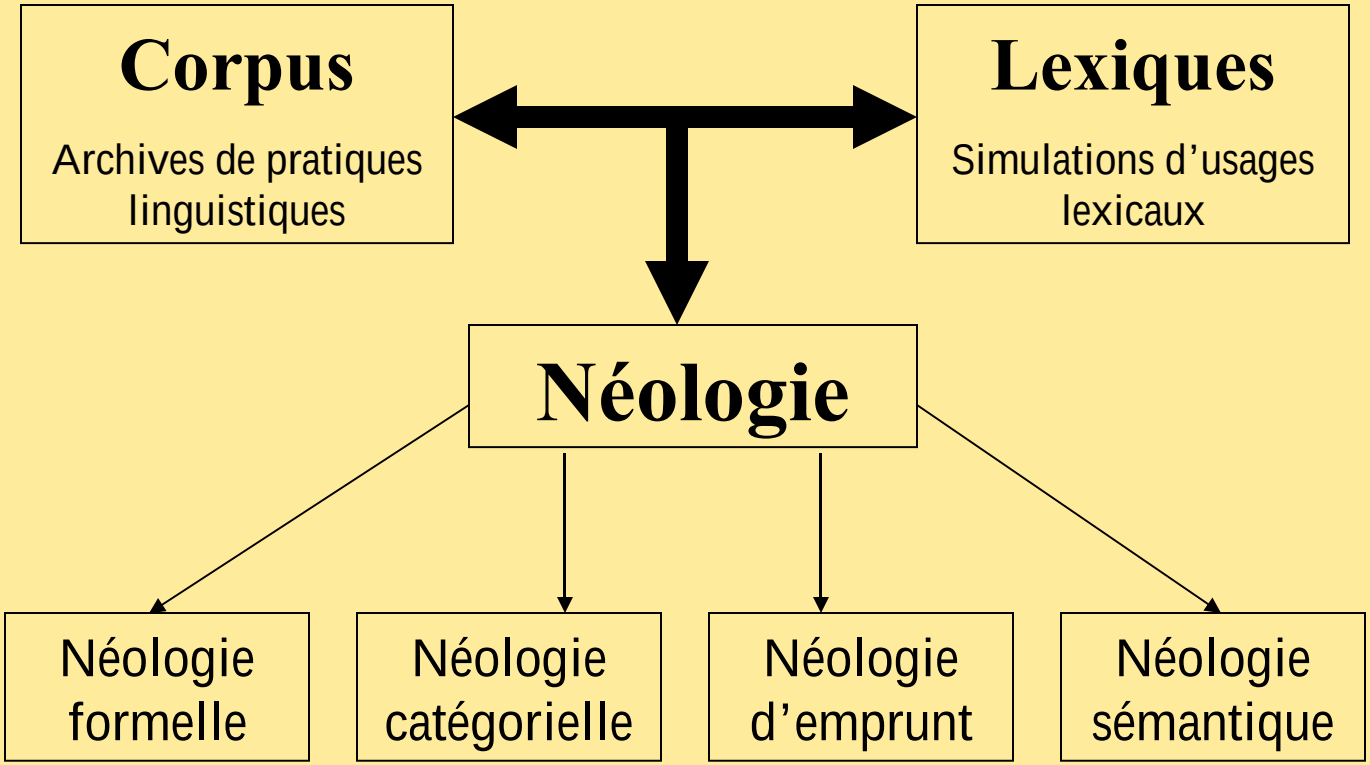
Contexte : la veille lexicale

- **Projet initié par S.Salmon-Alt, M.Valette et E.Petitjean**
- **Objectifs**
  - Exploitation et enrichissement des ressources de l'ATILF
  - Constitution de nouvelles ressources
  - Observation de la créativité lexicale
- **Adaptabilité**
  - odularité !"A#A$
  - %ormalisation &
    - Text Encodin ' Initiative !TEI$
    - Lexical ar(up Frame) or(!L F$
  - * i++usion !Interne , Externe & C% - TL$

La néologie

- **Néologie = Ensemble des unités lexicales nouvelles dans un état de langue donné**
- **Segmentation fiable en unités lexicales**
  - . ti / ueta ' e morphos0ntaxi / ue préalable des corpus
- **Comparaison à un état de langue antérieur**
  - 1 tilisation de lexi / ues d'exclusion
- **Appartenance à l'état de langue actuelle**
  - ultiplication des corpus d'observation
- **Prise en compte des spécificités des corpus**
  - T0polo ' ie2 date2 auteur

Méthodologie



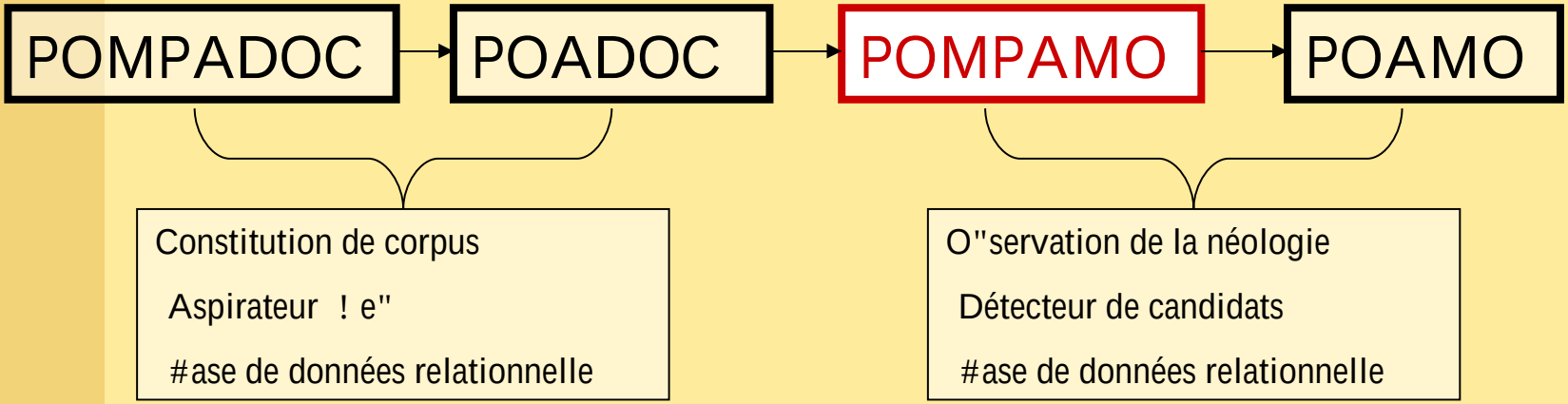
Les néologies (1)

- **Néologie formelle : formée par dérivation, composition, abréviation ou variation graphique**
  - Formes inconnues des lexi/ues
  - né'ationnisme2 médiatisation
- **Néologie catégorielle : Formes connues, mais sous une autre catégorie syntaxique**
  - 3 t0pes de dérivations détectés & %om Commun → Ad4ecti+ !ennemi$ et Ad4ecti+ → %om Commun !documentaire$
  - 5ensible 6 l'éti/ueta'e

Les néologies (2)

- **Néologie d'emprunt : unités lexicales empruntées à d'autres langues**
  - Formes inconnues des lexi/ues !sniper$
  - ...epérée? mais ne!distin'uée de la néolo'ie ormelle
- **Néologie sémantique : unités lexicales**

Plateforme de veille lexicale



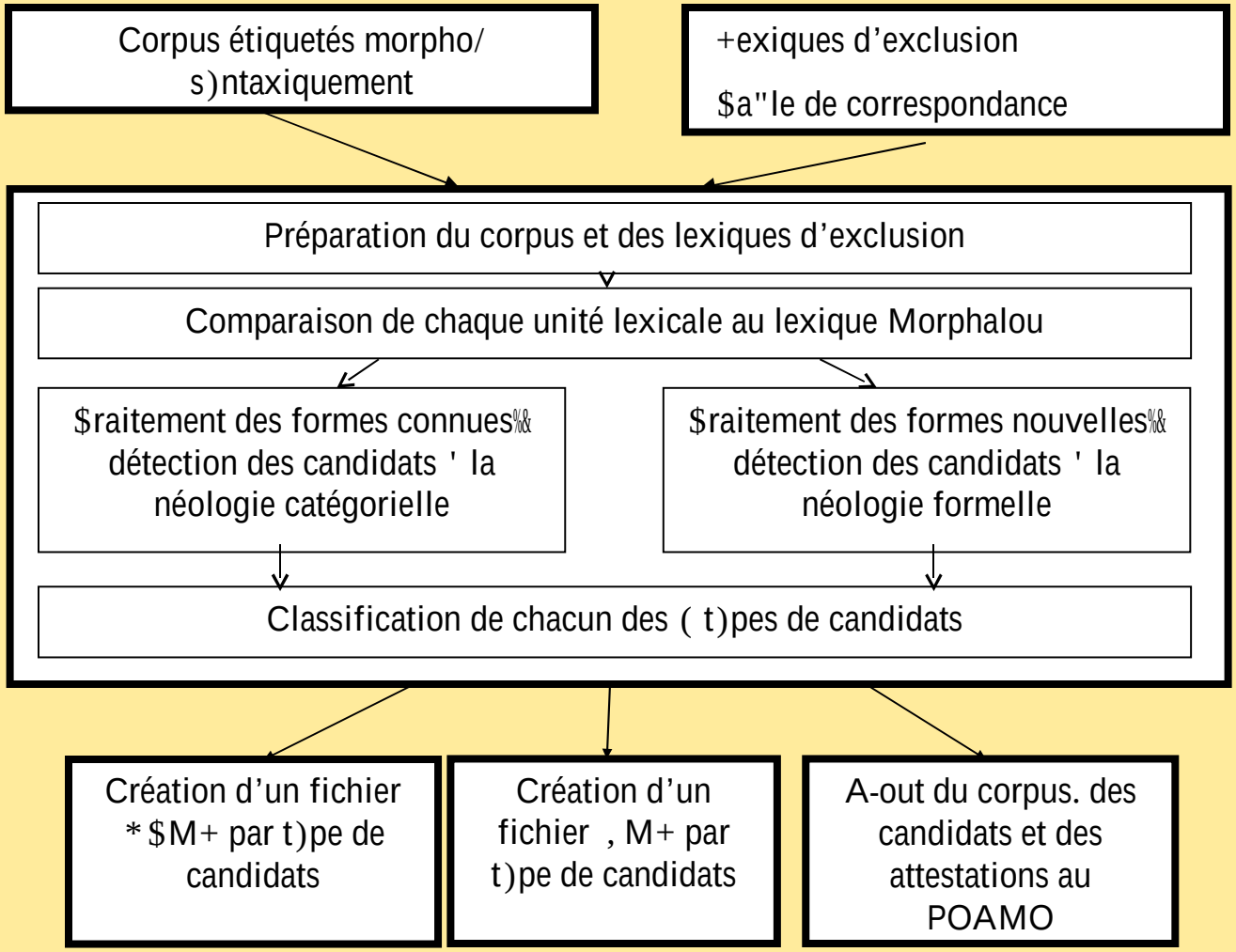
POMPADOC

- **Aspirateur de page Web**
  - - e / u9te par mots,cle+s
  - Interro ' ation de moteurs de recherche
- **État actuel**
  - : rotot0pe de "érém0 Ceintre0 et ; ori(: ete0
  - oteur de recherche & <oo ' le
  - : aramétra ' e par nb de mots2 position des mots, cle+s2 nom de domaine2 nb de pa ' es aspirées
  - Formats de sortie & =T L > 7 L ? TEI : @
- **A venir**
  - - ésoudre problAme encoda ' e 1TF,B
  - Enrichissement du +ormat 7 L
  - <énérer sortie T7T
  - Coupler avec : OA * OC

POADOC

- **Base de données de pages Web**
  - Case de donnée relationnelle
  - Interro ' ation croisée par méta,données !date2 t0pe de texte2 domaine2 ' enre2 auteur$
  - 5ortie & corpus
- **Réflexion ouverte sur les spécifications**
  - Calcul des +ré / uences D !nb de mots2 Ad42 %2 E $
  - * omanialisation D
  - Annotation morphos0ntaxi / ue D
  - Traitement sur textes en entrée ou corpus en sortieD
  - A4out d'un module de traitement supplémentaire pour ces enrichissementsD

POMPAMO


Corpus étiquetés morpho/ s)ntaxiquement
+exiques d'exclusion
$a"le de correspondance
Préparation du corpus et des lexiques d'exclusion
Comparaison de chaque unité lexicale au lexique Morphalou
$raitement des formes connues%& détection des candidats ' la néologie catégorielle
$raitement des formes nouvelles%& détection des candidats ' la néologie formelle
Classification de chacun des (t)pes de candidats
Création d'un fichier *$M+ par t)pe de candidats
Création d'un fichier , M+ par t)pe de candidats
A-out du corpus. des candidats et des attestations au POAMO

Données en entrée

- **Corpus**
  - 5e ' menté et éti / ueté
  - Format & sortie éti / ueteur ou 7 L TEI, : @
- **Options**
  - Taille des contextes d'attestation !max8 F@ phrases ou GHH mots$
  - Filtres pour candidats 6 la néolo ' ie +ormelle
    - Choix des lexi / ues
    - 5uppression des +ormes composées
    - 5uppression des +ormes éti / uetées % :
- **Table de correspondance**
  - . ti / uettes propriétaires? éléments et attributs standards L F,I5O TC GI 5CJ

Lexiques d’exclusions

- **Lexique principal de formes fléchies du français : MORPHALOU 2.0**
- **Validité linguistique (Nomenclature TLF)**
  - Lar ' e couverture !@3J I3@ ⁺ormes

LexiqueU.0bvgi: :anêe

LexiqueU.0bvgi: :anêe

Préparation et Comparaison (1)

- **Préparation**
  - Anal0se du Corpus & récupération des unités lexicales
  - Optimisation de l'accAs aux ressources par la création de sous,lexi / ues
- **Comparaison des unités lexicales du corpus au lexique principal**
  - * istinction de t0pes &

    formes connues & potentielle néologie catégorielle

    formes nouvelles & potentielle néologie formelle

Préparation et Comparaison (2)

Corpus‰
+e négationnisme . une "ar"arie "analisée 0+e 1igaro. (2342 3(4445 6osovo

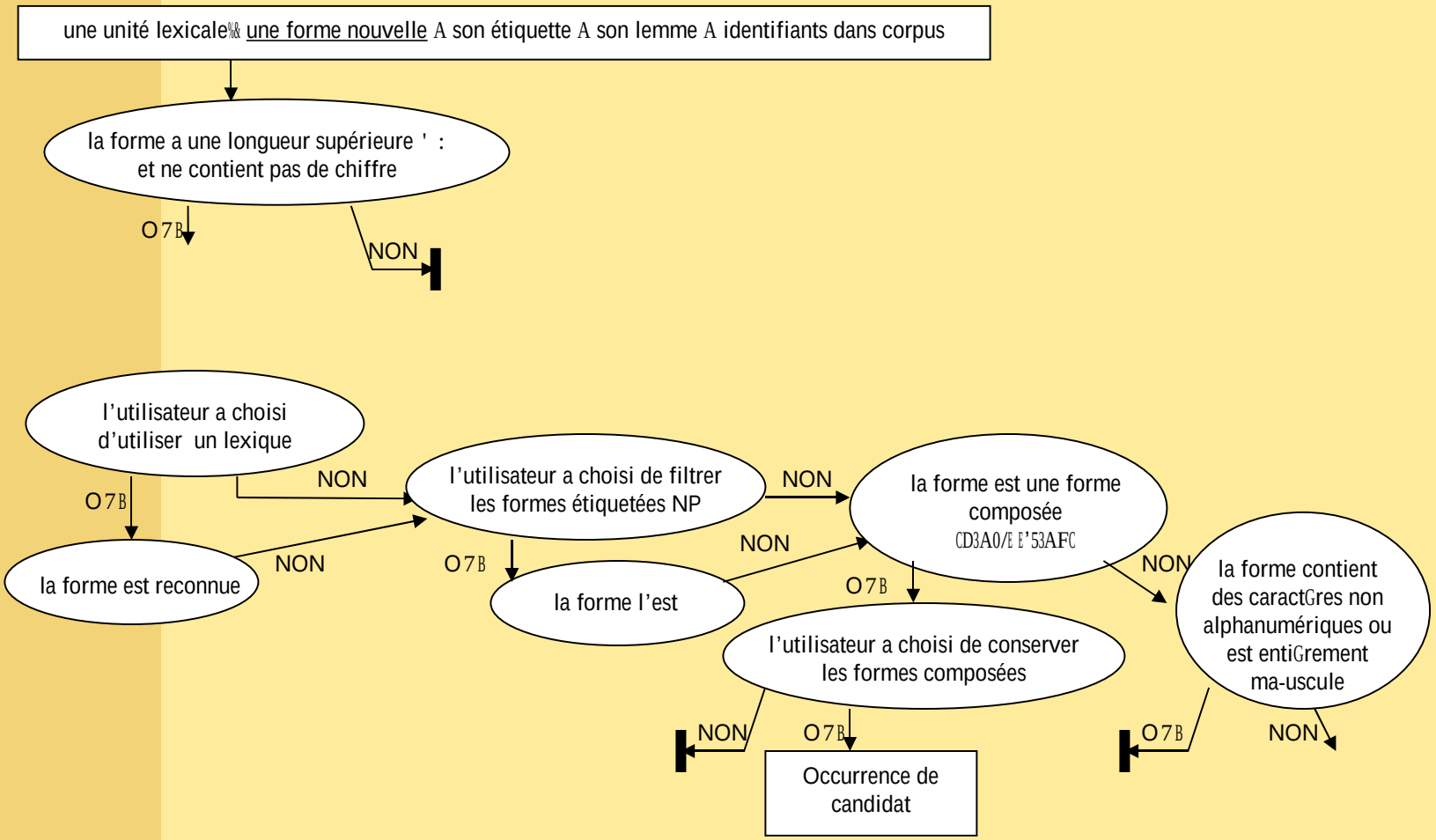
7nités lexicales‰
+e Da/ms/d 89:;<< le
négationnisme Ncms 89:;<= négationnisme
. >p8 89:;
0 ? 5
"ararae

Traitement des formes connues

- **Traitement des formes étiquetées Adj. et Nc**
- **La forme est-elle répertoriée sous cette cat. grammaticale dans le lexique principal?**
  - O1I N pas de néolo'ie
  - %O% N poursuite du traitement
- **La forme est-elle répertoriée sous la 2nd cat. grammaticale dans le lexique principal?**
  - O1I N néolo'ie caté'orielle
  - %O% N pas de néolo'ie
- **Exemple**
  - OPdocumentaire2 %cms2 documentairePQ
  - - épertorié comme %c dans orphalou N %O%
  - - épertorié comme Ad48 dans orphalou N O1I

  N Candidat 6 la néolo'ie caté'orielle

Traitement des formes nouvelles



Regroupement en candidats

- **Création d'un tableau de candidats :**
- **Forme + étiquette + lemme**
  - Ensemble d'attestations !localisation$

négationnisme Ncms négationnisme

89H:= sentence9<4 paragraph9((H4I H:I H:<
89:=:: sentence9=(paragraph9(2 :=42 :=:2 :=:4
89:;<= sentence9;4 paragraph9<: :;(I :;<I :;<<
89(<=I sentence9I< paragraph9=4 (<=((<2((<=@

- +imite gauche du contexte d'attestation
- +imite droite du contexte d'attestation
- +ocalisation du candidat en n" d'unités lexicales

Création d'un fichier HTML par type de candidats

candidats à la néologie catégorielle

Candidats					
	orthographe ▼▲	hypothèse de lemmatisation ▼▲	hypothèse d'analyse morphosyntaxique ▼▲	fréquence absolue ▼▲	fréquence au sein du corpus (%) ▼▲
1	dissident	dissident	Ncms	1	20.00
2	documentaire	documentaire	Ncms	1	20.00
3	médiatique	médiatique	Afp.s	2	40.00
4	tout	tout	Afpms	1	20.00

Attestations	
	extrait(s)
1	1. Quant à Télérama, si le magazine a certes accordé dans son numéro du 7 mai 2003 une interview au dissident, il n'a pas à ma connaissance publié dans ses colonnes de critiques de ses ouvrages ni même, curieusement, du documentaire sorti en septembre dernier Noam Chomsky :
2	1. Quant à Télérama, si le magazine a certes accordé dans son numéro du 7 mai 2003 une interview au dissident, il n'a pas à ma connaissance publié dans ses colonnes de critiques de ses ouvrages ni même, curieusement, du documentaire sorti en septembre dernier Noam Chomsky :
3	1. Une partie importante de son travail est consacrée à établir les preuves objectives de l'existence d'une propagande médiatique.
4	1. Il y apprend que tout le monde fait quelque chose d'important.

Export XML TEI-LMF

```
<lexicalEntry id="entry_45">
  <formSet>
    <lemmatizedForm processStatus="provisionallyProcessed">
      <orthography>négationnisme</orthography>
      <grammaticalCategory>commonNoun</grammaticalCategory>
      <grammaticalGender>masculine</grammaticalGender>
    </lemmatizedForm>
    <inflectedForm>
      <orthography>négationnisme</orthography>
      <grammaticalNumber processStatus="provisionallyProcessed">singular</grammaticalNumber>
    </inflectedForm>
  </formSet>
  <sense>
    <dicteg>
      <cit id="cit_45_1">
        <q> Enfin, les accusations de<oRef>négationnisme</oRef> trouvent leurs source dans</q>
        <bibl>
          <ref word_id="w_914" sentence_id="sentence_30" paragraph_id="paragraph_22"/>
        </bibl>
      </cit>
      <cit id="cit_45_2">
        <q> 1998, il décrivait le<oRef>négationnisme</oRef> comme la pire atrocité</q>
        <bibl>
          <ref word_id="w_1411" sentence_id="sentence_42" paragraph_id="paragraph_25"/>
        </bibl>
      </cit>
    </dicteg>
  </sense>
</lexicalEntry>
```

Implémentation

- **Langage de programmation**
  - "avaT 3 : lat+orm 5tandard Edition @8H
- **Algorithmique**
  - Al ' orithmes de tri dichotomi / ue
  - A : I 5A7
  - AccAs base de données
- **Bases de données**
  - 05RL
- **Documents semi-structurés / Standards**
  - 7 L2 75LT2 =T L
  - TEI2 L F
- **Etiqueteur morpho-syntaxique**
  - Cordial Anal0seur

Perspectives Ressources

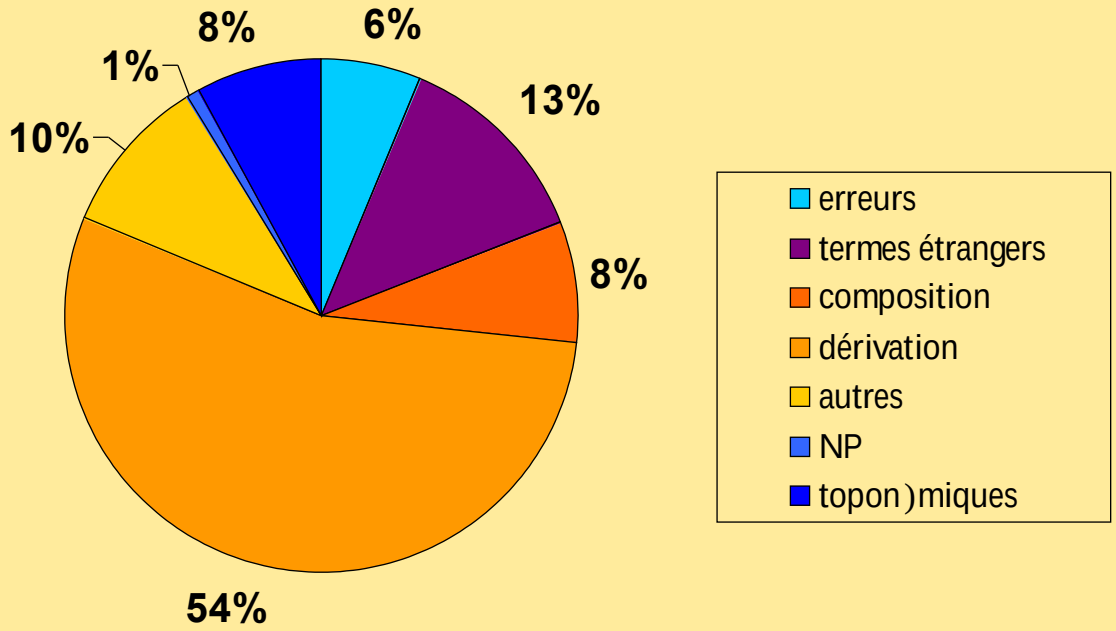
- **Lexiques**
  - Ac/uisitions nouvelles !% : 25i ' les et acron0mes$
  - Format standard
  - Choix du lexi/ue principal
- **Corpus en entrée**
  - * iversi+ication des +ormats
  - * iversi+ication des éti/ueteurs
- **Diffusion**
  - Création d'une inter+ace ' raphi/ue
  - ise en li ' ne sur le site du C% - TL !))) 8cnrtl8+r$

Évaluation

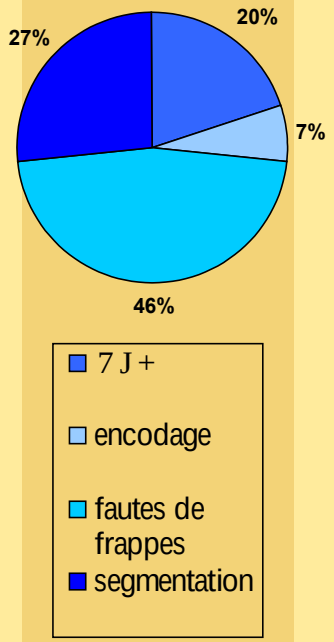
- **Données textuelles :**
  - *« Le Monde diplomatique »*
  - *Année 1998*
  - *Auteurs multiples*
- **Type de texte :**
  - *Discours journalistique*
  - *Genre majoritaire : article*
  - *Domaine majoritaire : géopolitique*
-
  - *2119 candidats à la néologie formelle*
  - *312 candidats à la néologie catégorielle*
- **Temps d'exécution : 125 secondes**

Candidats à la néologie formelle

- **264 candidats commençant par la lettre A, pour 477 occurrences**

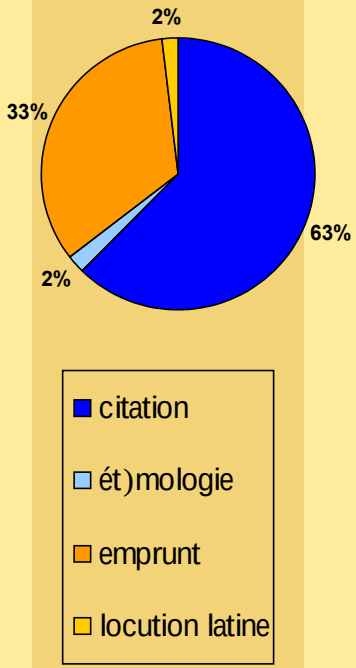


Erreurs



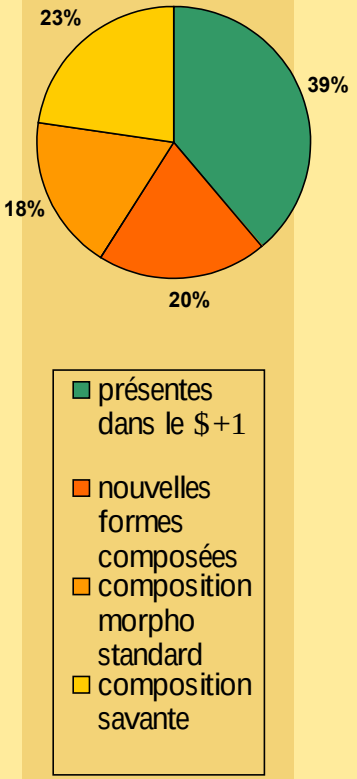
- **Formes appartenant à des adresses de sites Internet (3 formes, 3 occurrences) : acdi-cida**
- **Formes issues d'un mauvais traitement de l'encodage de caractères (1 forme, 4 occurrences): amp**
- **Fautes de frappes (7 formes, 7 occurrences): annnés**
- **Erreurs de segmentation (4 formes, 8 occurrences): au-boutistes pour jusqu'au-boutistes**

Termes étrangers



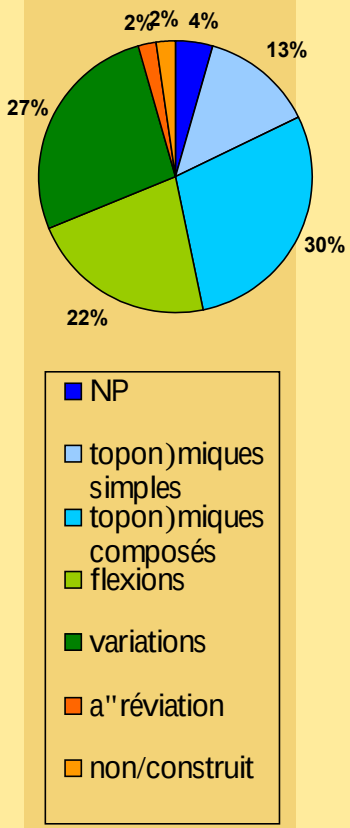
- **En contexte de citation (13 formes, 26 occurrences) : « ce que l'ancien ministre (…) désigne, en portugais, par une aculturaçao europeia. »**
- **En contexte étymologique (1 forme, 1 occurrence) : « Que signifie autonome ? Cela veut dire autosnomos, qui se donne à soi -même sa loi. »**
- **En contexte d'emprunt (16 formes, 41 occurrences) : « sur fond de discorde entre juifs et Arabes, ashkénazes et orientaux, laïcs et religieux, riches et pauvres... »**
- **Locution latine (1 forme, 1 occurrence) : ad vitam aeternam**

Locutions, Formes Composées et Composition Morphologique



- **Locutions et formes composées présentes dans le TLF mais absentes de Morphalou (17 formes, 99 occurrences): à l'encontre**
- **Nouvelles formes composées, figement à partir de combinaisons syntaxiques (9 formes, 12 occurrences): assurance-chômage**
- **Composition morphologique standard (8 formes, 8 occurrences): anarcho-syndicalisme**
- **Composition savante (10 formes, 14 occurrences): agrofournisseurs**

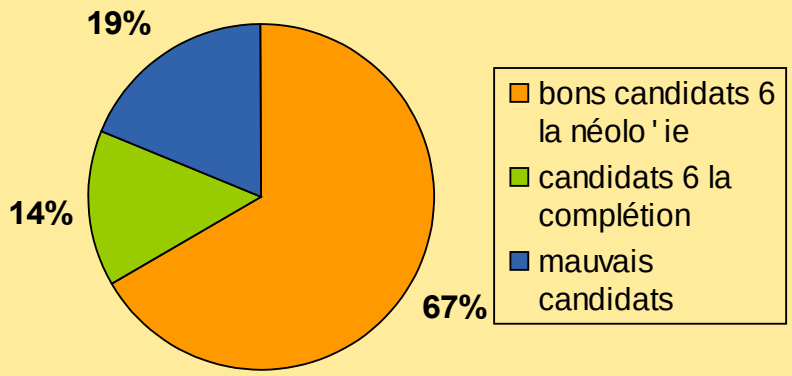
NP, toponymiques et autres



- **Noms Propres non reconnus par l'étiqueteur (2 formes, 2 occurrences): arrap Moi**
- **Noms et adjectifs toponymiques simples (6 formes, 6 occurrences): alavaise**
- **Noms et adjectifs toponymiques composés (13 formes, 18 occurrences): argentino-brésiliens**
- **Formes fléchies de lemmes répertoriés dans Morphalou (10 formes, 14 occurrences): arrière-pensées**
- **Variations graphiques de formes répertoriées dans Morphalou (11 formes, 33 occurrences): autodéfense**
- **Abréviation (1 forme, 1 occurrence): amphi**
- **Unité lexicale non construite (1 forme, 1 occurrence): auteure**

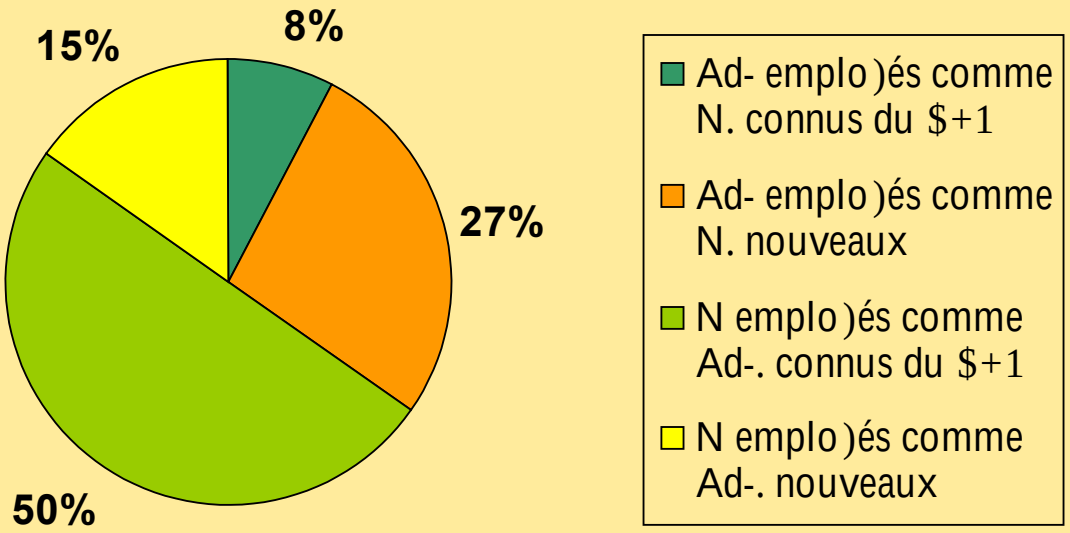
Bilan néologie formelle

- **175 bons candidats, associés à 241 contextes d'attestation,dont l'observation dans le POAMO permettra d'évaluer la « vitalité » en fonction des genres, types, domaines, auteurs et périodes.**
- **38 candidats à la complétion directe de Morphalou, associés à 146 attestations**
- **50 mauvais candidats, associés à 75 contextes d'attestation, dont l'observation peut permettre une diminution du bruit**

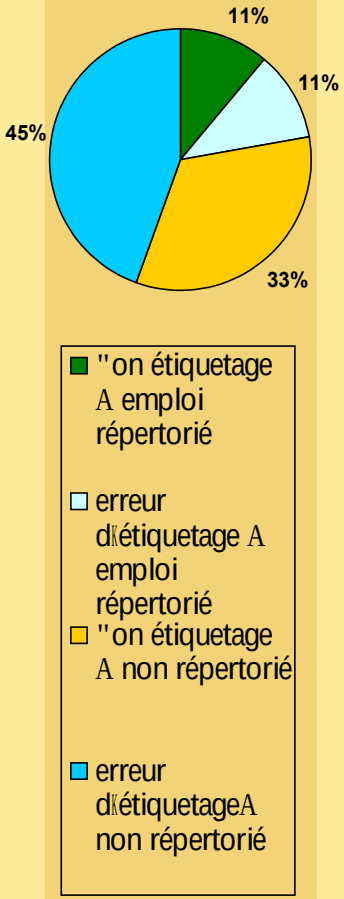


Candidats à la néologie catégorielle

- **26 candidats commençant par la lettre A, pour 51 occurrences**
- **1) consultation du TLF**
- **2) vérification de l'étiquetage, en contexte**

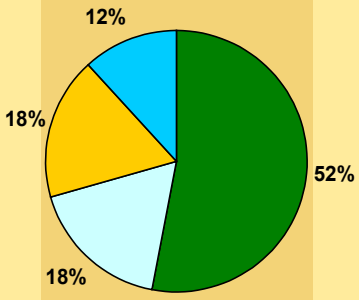


Adjectifs étiquetés Substantif



	Emploi répertorié dans le TLF	Emploi non répertorié dans le TLF
. ti / ueta ' e correct	F +orme2F occurrence OPles Alsaciens expriment lSespoir de ! E $Q	G +ormes2 F@ occurrences O les LHH HHH autochtones licenciés en FKK I PQ
. ti / ueta ' e incorrect	F +orme2F occurrence OPles di++érents appels 6 connotation anti4uivePQ	J +ormes2 J occurrences OPla périphérie2 atomisée2 désordonnéePQ

Substantifs étiquetés adjectifs



	Emploi répertorié dans le TLF	Emploi non répertorié dans le TLF
. ti / ueta ' e correct	K +ormes2 3B occurrences OP / ui a été habituée 6 la variété an ' lo, saxonnePQ	G +ormes2 @ occurrences OPtous les noms de cito0ens américains ou amis des Etats, 1 nisPQ
. ti / ueta ' e incorrect	G +ormes2J occurrences OPar ' ent ' aspillé en éléphants blancs PQ	3 +ormes2 L occurrences O / ui n$ont cessé2 avant comme aprAs la colonisation Q

Bilan néologie catégo

Perspective veille : POAMO

- **Observatoire de créativité lexicale**
  - Case de données relationnelle
  - Entrée & sortie de : O : A O
- **Interrogations croisées**
  - - e / u9tes sur méta,données
  - - e / u9tes sur ŧormes et expressions ré ' uliAres
  - Calculs de ŧré / uences
- **Caractérisation des candidats et sélection**
  - Enrichissement lexicométri / ue
  - . volution diachroni / ue
  - - épartition entre t0pes de corpus