ANALYSE ET TRAITEMENT
INFORMATIQUE
DE LA LANGUE FRANÇAISE

Unité mixte de recherche 7118

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

NANCY 2 UNIVERSITE

Université HENRI POINCARÉ NANCY 1

POMPAMO :
Détection automatique de candidats à la néologie

Sandrine Ollinger
sandrine.ollinger@atilf.fr

http://www.atilf.fr

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**
  - Modularité (JAVA)
  - Normalisation :
    - Text Encoding Initiative (TEI)
    - Lexical Markup Framework (LMF)
  - Diffusion (Interne - Externe : CNRTL)

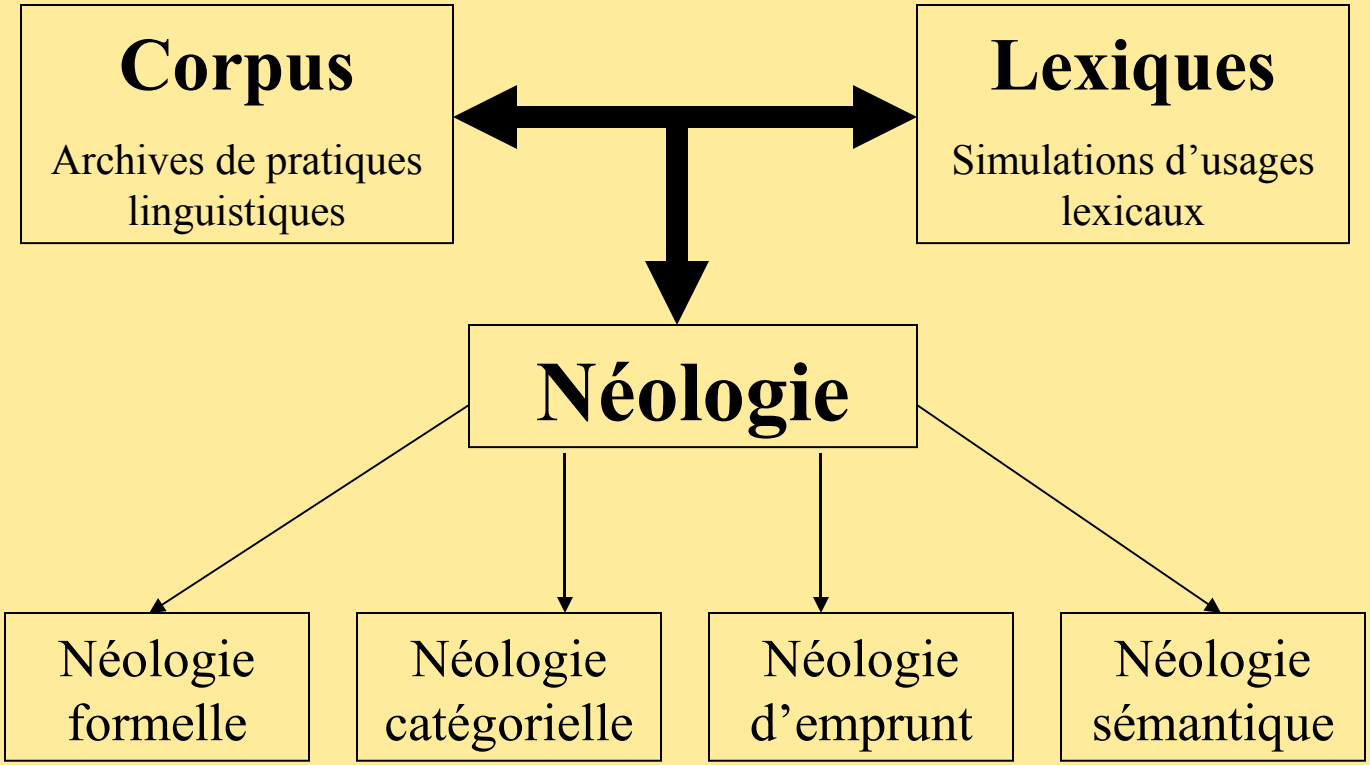
sandrine.ollinger@atilf.fr

La néologie

- **Néologie = Ensemble des unités lexicales nouvelles dans un état de langue donné**
- **Segmentation fiable en unités lexicales**
  - Étiquetage morphosyntaxique préalable des corpus
- **Comparaison à un état de langue antérieur**
  - Utilisation de lexiques d'exclusion
- **Appartenance à l'état de langue actuelle**
  - Multiplication des corpus d'observation
- **Prise en compte des spécificités des corpus**
  - Typologie, date, auteur

sandrine.ollinger@atilf.fr

Méthodologie



sandrine.ollinger@atilf.fr

Les néologies (1)

- **Néologie formelle : formée par dérivation, composition, abréviation ou variation graphique**
  - Formes inconnues des lexiques
  - négationnisme, médiatisation

- **Néologie catégorielle : Formes connues, mais sous une autre catégorie syntaxique**
  - 2 types de dérivations détectés : Nom Commun → Adjectif (ennemi) et Adjectif → Nom Commun (documentaire)
  - Sensible à l'étiquetage

sandrine.ollinger@atilf.fr

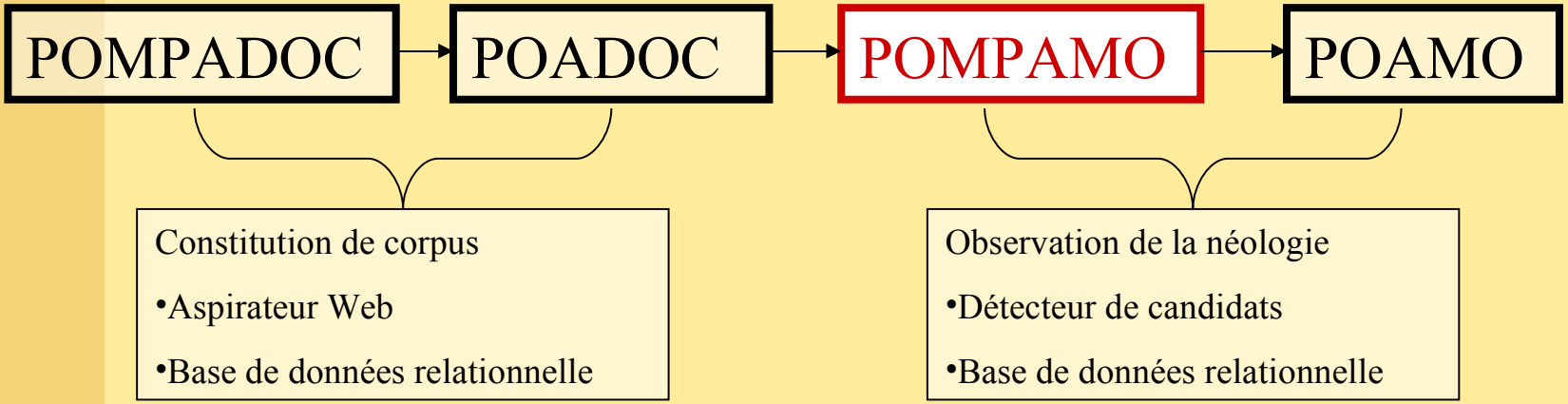
Les néologies (2)

- **Néologie d'emprunt : unités lexicales empruntées à d'autres langues**
  - Formes inconnues des lexiques (snipers)
  - Repérée, mais non distinguée de la néologie formelle

- **Néologie sémantique : unités lexicales ayant subi une extension, une restriction ou un changement complet de sens**
  - Formes connues, sans changements morphosyntaxiques
  - Non détectée (souris)
  - Travaux en cours : DIXEM (E.Jacquey et M.Valette)

sandrine.ollinger@atilf.fr

Plateforme de veille lexicale



sandrine.ollinger@atilf.fr

POMPADOC

- **Aspirateur de page Web**
  - Requête par mots-clefs
  - Interrogation de moteurs de recherche
- **État actuel**
  - Prototype de Jérémy Ceintrey et Yorik Petey
  - Moteur de recherche : Google
  - Paramétrage par nb de mots, position des mots-clefs, nom de domaine, nb de pages aspirées
  - Formats de sortie : HTML + XML / TEI P5
- **A venir**
  - Résoudre problème encodage UTF-8
  - Enrichissement du format XML
  - Générer sortie TXT
  - Coupler avec POADOC

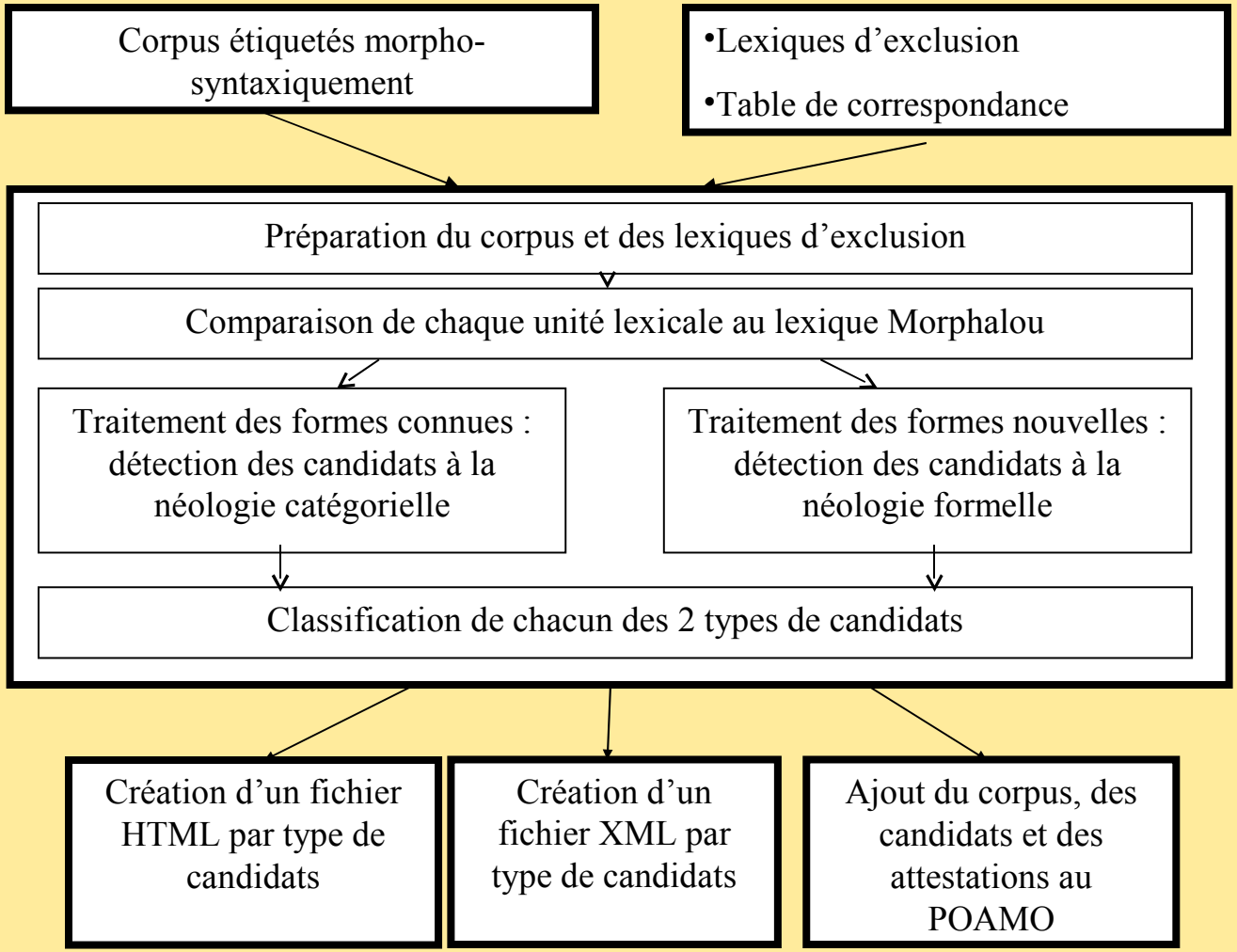
sandrine.ollinger@atilf.fr

POADOC

- **Base de données de pages Web**
  - Base de donnée relationnelle
  - Interrogation croisée par méta-données (date, type de texte, domaine, genre, auteur)
  - Sortie : corpus
- **Réflexion ouverte sur les spécifications**
  - Calcul des fréquences ? (nb de mots, Adj, N,…)
  - Domanialisation ?
  - Annotation morphosyntaxique ?
  - Traitement sur textes en entrée ou corpus en sortie?
  - Ajout d'un module de traitement supplémentaire pour ces enrichissements?

sandrine.ollinger@atilf.fr

POMPAMO



sandrine.ollinger@atilf.fr

Données en entrée

- **Corpus**
  - Segmenté et étiqueté
  - Format : sortie étiqueteur ou XML TEI-P5
- **Options**
  - Taille des contextes d'attestation (max. 15 phrases ou 300 mots)
  - Filtres pour candidats à la néologie formelle
    - Choix des lexiques
    - Suppression des formes composées
    - Suppression des formes étiquetées NP
- **Table de correspondance**
  - Étiquettes propriétaires/ éléments et attributs standards LMF-ISO TC 37 SC4

sandrine.ollinger@atilf.fr

Lexiques d'exclusions

- **Lexique principal de formes fléchies du français : MORPHALOU 2.0**
- **Validité linguistique (Nomenclature TLF)**
  - Large couverture (524 725 formes fléchies 95 810 lemmes)
  - Accès libre au format XML - LMF
- **Lexiques supplémentaires inclus**
  - 70 438 Noms propres (ABU, Prolex, Tagen)
  - 6 903 Adjectifs toponymiques et gentilés (Prolex)
    (le vin français, les Français)
  - 140 nombres composés
- **Lexiques supplémentaires utilisateur**

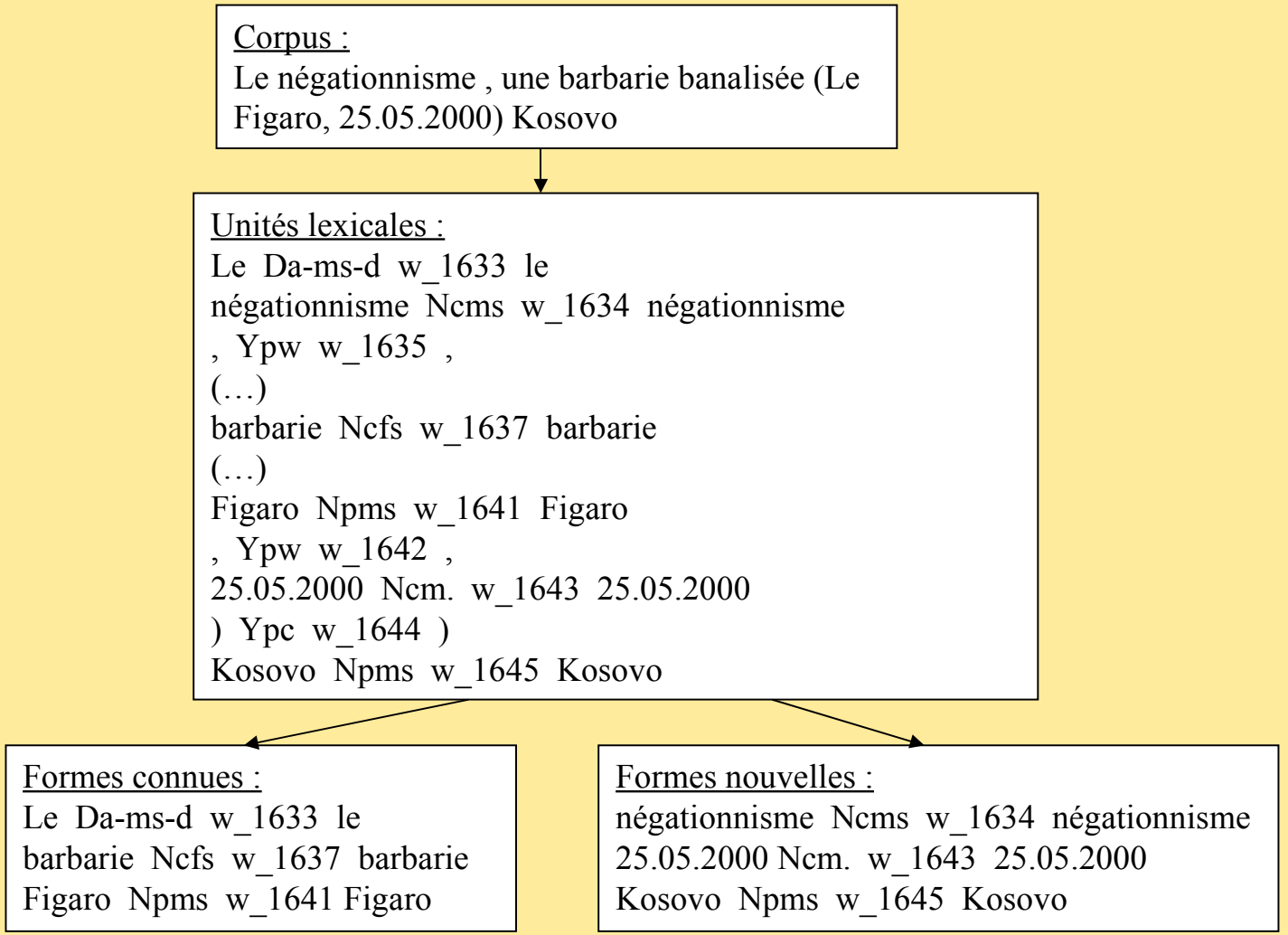
sandrine.ollinger@atilf.fr

Préparation et Comparaison (1)

- **Préparation**
  - Analyse du Corpus : récupération des unités lexicales
  - Optimisation de l'accès aux ressources par la création de sous-lexiques
- **Comparaison des unités lexicales du corpus au lexique principal**
  - Distinction de types :
    - formes connues : potentielle néologie catégorielle
    - formes nouvelles : potentielle néologie formelle

sandrine.ollinger@atilf.fr

Préparation et Comparaison (2)



sandrine.ollinger@atilf.fr

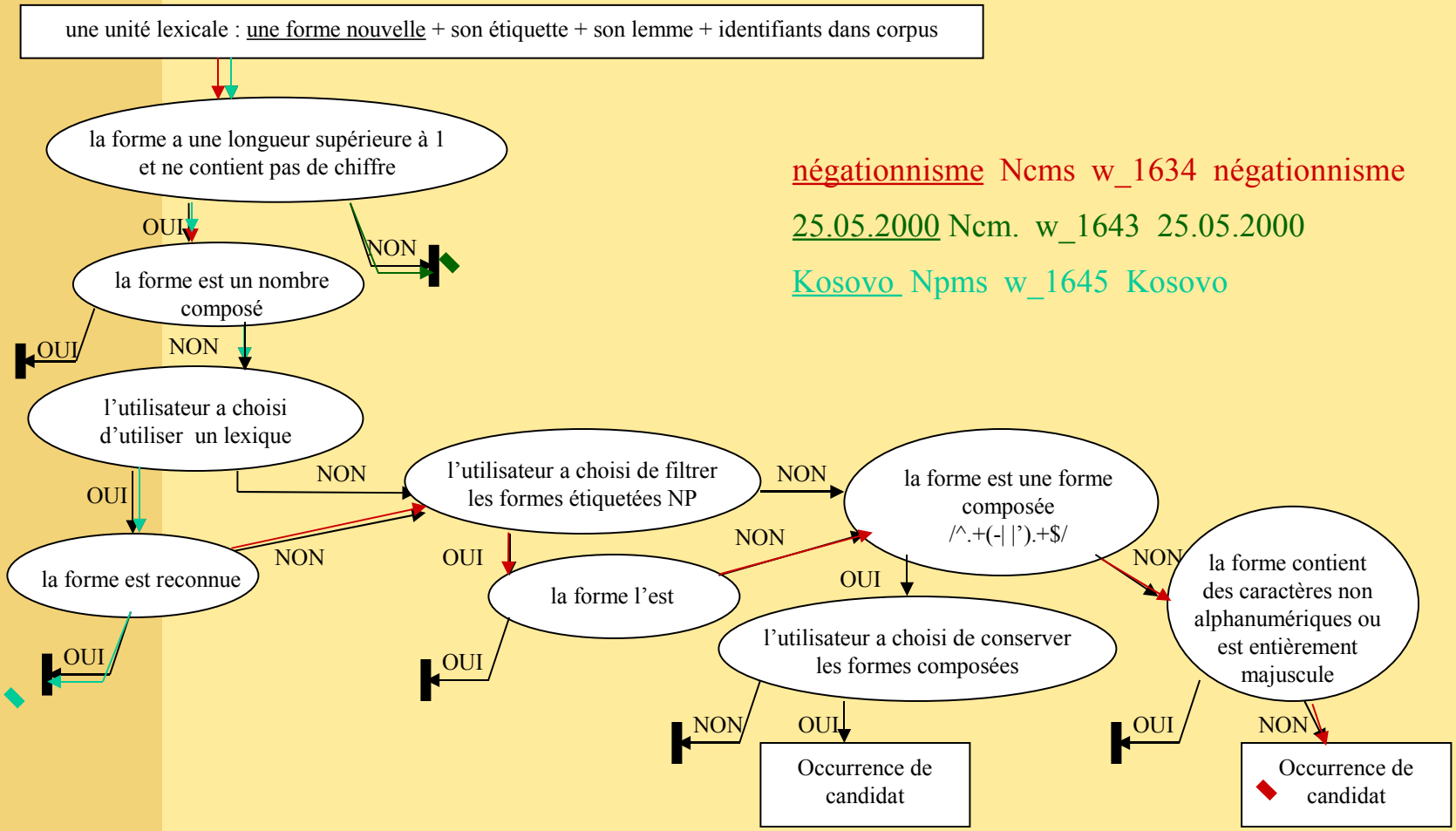
Traitement des formes connues

- **Traitement des formes étiquetées Adj. et Nc**
- **La forme est-elle répertoriée sous cette cat. grammaticale dans le lexique principal?**
  - OUI → pas de néologie
  - NON → poursuite du traitement
- **La forme est-elle répertoriée sous la 2nd cat. grammaticale dans le lexique principal?**
  - OUI → néologie catégorielle
  - NON → pas de néologie
- **Exemple**
  - « documentaire, Ncms, documentaire »
  - Répertorié comme Nc dans Morphalou → NON
  - Répertorié comme Adj. dans Morphalou → OUI

  → Candidat à la néologie catégorielle

sandrine.ollinger@atilf.fr

Traitement des formes nouvelles



négationnisme Ncms w_1634 négationnisme

25.05.2000 Ncm. w_1643 25.05.2000

Kosovo Npms w_1645 Kosovo

sandrine.ollinger@atilf.fr

Regroupement en candidats

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

négationnisme Ncms négationnisme

w_914 sentence_30 paragraph_22 908 918 913
w_1411 sentence_42 paragraph_25 1405 1415 1410
w_1634 sentence_60 paragraph_31 1628 1638 1633
w_2348 sentence_83 paragraph_40 2342 2352 2347

- Limite gauche du contexte d'attestation
- Limite droite du contexte d'attestation
- Localisation du candidat en nb d'unités lexicales

sandrine.ollinger@atilf.fr

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. 2. et D. Schneidermann - Le cauchemar médiatique - Denoel, 2003, p 121-122
4	1. Il y apprend que tout le monde fait quelque chose d'important.

sandrine.ollinger@atilf.fr

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>
```


sandrine.ollinger@atilf.fr

Implémentation

- **Langage de programmation**
  - JavaTM 2 Platform Standard Edition 5.0
- **Algorithmique**
  - Algorithmes de tri dichotomique
  - API SAX
  - Accès base de données
- **Bases de données**
  - MySQL
- **Documents semi-structurés / Standards**
  - XML, XSLT, HTML
  - TEI, LMF
- **Etiqueteur morpho-syntaxique**
  - Cordial Analyseur

sandrine.ollinger@atilf.fr

Perspectives Ressources

- **Lexiques**
  - Acquisitions nouvelles (NP,Sigles et acronymes)
  - Format standard
  - Choix du lexique principal
- **Corpus en entrée**
  - Diversification des formats
  - Diversification des étiqueteurs
- **Diffusion**
  - Création d'une interface graphique
  - Mise en ligne sur le site du CNRTL (www.cnrtl.fr)

sandrine.ollinger@atilf.fr

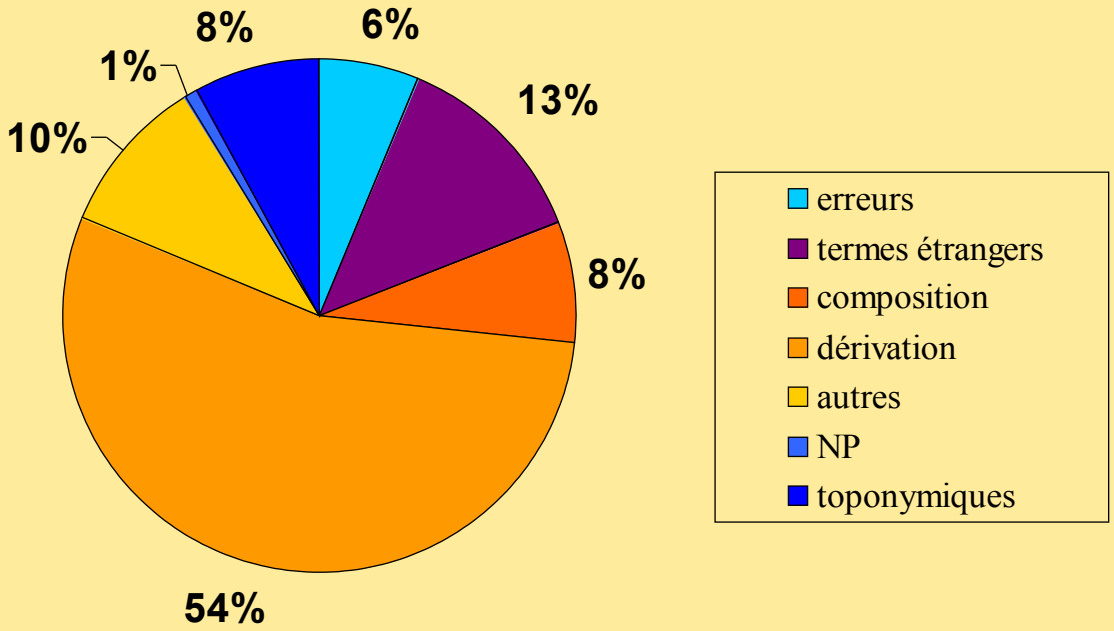
Évaluation

- **Données textuelles :**
  - *« Le Monde diplomatique »*
  - *Année 1998*
  - *Auteurs multiples*
- **Type de texte :**
  - *Discours journalistique*
  - *Genre majoritaire : article*
  - *Domaine majoritaire : géopolitique*
- ***501 691 unités lexicales, 19527 phrases***
  - *2119 candidats à la néologie formelle*
  - *312 candidats à la néologie catégorielle*
- **Temps d'exécution : 125 secondes**

sandrine.ollinger@atilf.fr

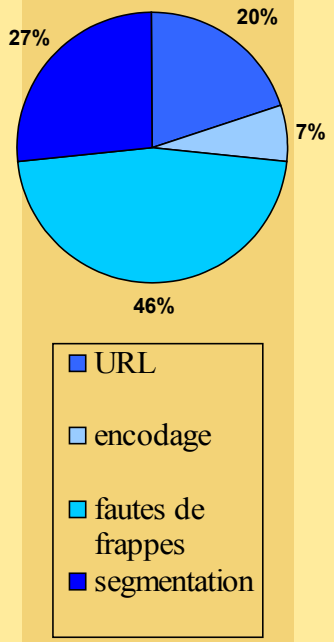
Candidats à la néologie formelle

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



sandrine.ollinger@atilf.fr

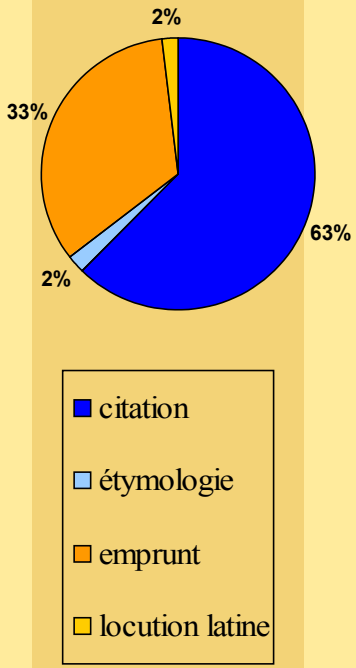
Erreurs



- **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**

sandrine.ollinger@atilf.fr

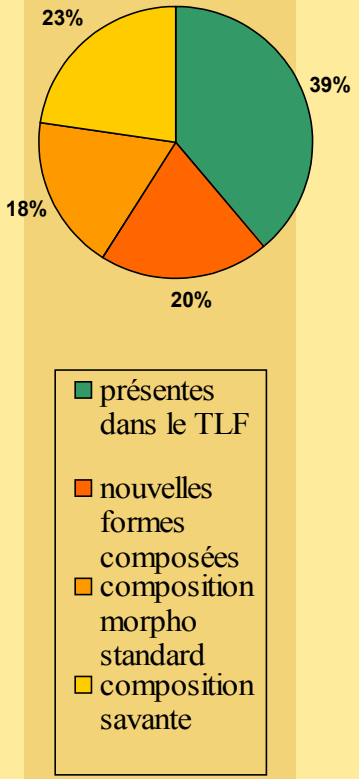
Termes étrangers



- **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**

sandrine.ollinger@atilf.fr

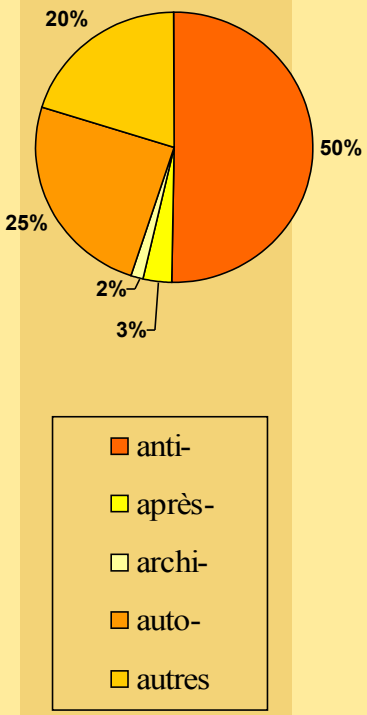
Locutions, Formes Composées et Composition Morphologique



- **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**

sandrine.ollinger@atilf.fr

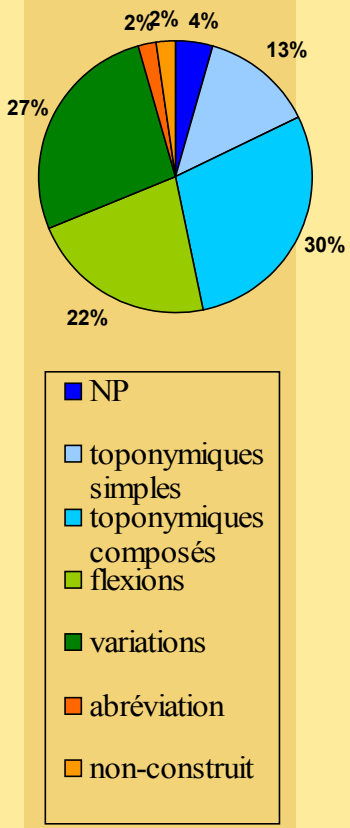
Dérivation



- **Candidats formés à l'aide du préfixe anti- (65 formes, 86 occurrences): anti-étatique**
- **Candidats formés à l'aide du préfixe auto- (32 formes, 35 occurrences): autocensure**
- **Candidats formés à l'aide du préfixe après- (4 formes, 7 occurrences): après-Lomé**
- **Candidats formés à l'aide du préfixe archi- (2 formes, 2 occurrences): archiminoritaires**
- **Candidats formés à l'aide d'autres affixes (26 formes, 33 occurrences) : autonomiser**

sandrine.ollinger@atilf.fr

NP, toponymiques et autres

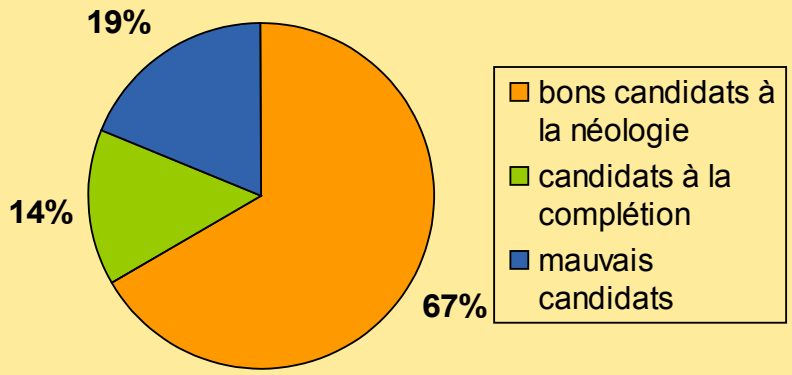


- **Noms Propres non reconnus par l'étiqueteur (2 formes, 2 occurrences): arrap Moi**
- **Noms et adjectifs toponymiques simples (6 formes, 6 occurrences): alavaise**
- **Noms et adjectifs toponymiques composés (13 formes, 18 occurrences): argentino-brésiliens**
- **Formes fléchies de lemmes répertoriés 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**

sandrine.ollinger@atilf.fr

Bilan néologie formelle

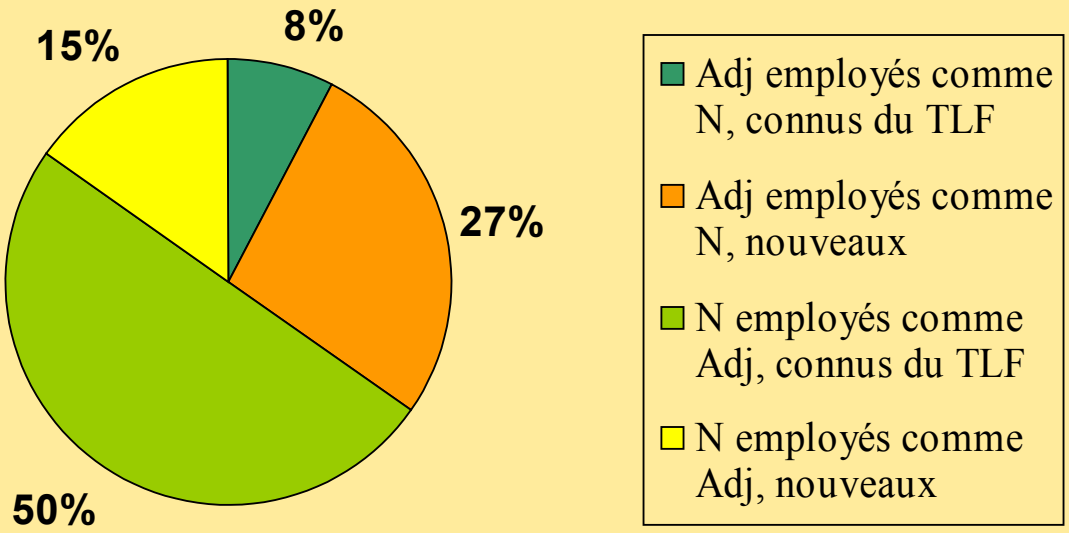
- **175 bons candidats, associés à 241 contextes d'attestation,dont l'observation dans le POAMO permettra d'évaluer la « vitalité » en fonction des genres, types, domaines, auteurs et périodes.**
- **38 candidats à la complétion directe de Morphalou, associés à 146 attestations**
- **50 mauvais candidats,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



sandrine.ollinger@atilf.fr

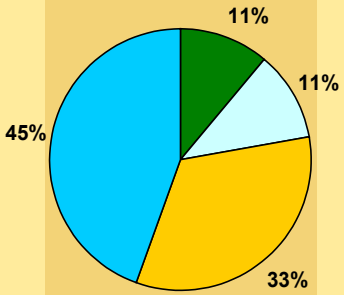
Candidats à la néologie catégorielle

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



sandrine.ollinger@atilf.fr

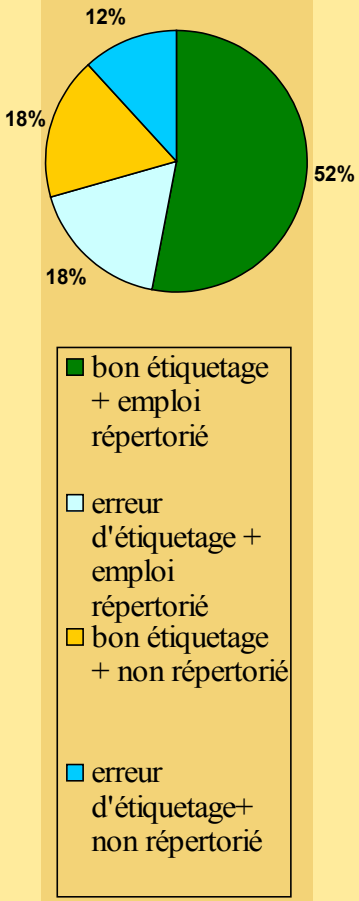
Adjectifs étiquetés Substantif



	Emploi répertorié dans le TLF	Emploi non répertorié dans le TLF
Étiquetage correct	1 forme,1 occurrence « les Alsaciens expriment l'espoir de (…)»	3 formes, 15 occurrences « les 600 000 autochtones licenciés en 1997 »
Étiquetage incorrect	1 forme,1 occurrence « les différents appels à connotation antijuive »	4 formes, 4 occurrences « la périphérie, atomisée, désordonnée »

sandrine.ollinger@atilf.fr

Substantifs étiquetés adjectifs

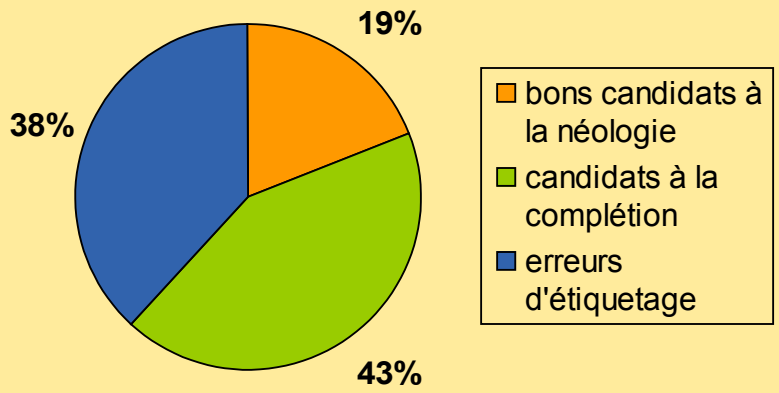


	Emploi répertorié dans le TLF	Emploi non répertorié dans le TLF
Étiquetage correct	9 formes, 28 occurrences « qui a été habituée à la variété anglo-saxonne »	3 formes, 5 occurrences « tous les noms de citoyens américains ou amis des Etats-Unis »
Étiquetage incorrect	3 formes,4 occurrences « argent gaspillé en éléphants blancs »	2 formes, 6 occurrences «qui n'ont cessé, avant comme après la colonisation »

sandrine.ollinger@atilf.fr

Bilan néologie catégorielle

- **5 bons candidats, associés à 19 contextes d'attestation,dont l'observation dans le POAMO permettra d'évaluer la « vitalité » en fonction des genres, types, domaines, auteurs et périodes.**
- **11 candidats à la complétion directe de Morphalou, associés à 30 attestations**
- **10 erreurs d'étiquetage, associés à 15 contextes d'attestation, dont l'observation peut permettre une diminution du bruit**



sandrine.ollinger@atilf.fr

Perspective veille : POAMO

- **Observatoire de créativité lexicale**
  - Base de données relationnelle
  - Entrée : sortie de POMPAMO
- **Interrogations croisées**
  - Requêtes sur méta-données
  - Requêtes sur formes et expressions régulières
  - Calculs de fréquences
- **Caractérisation des candidats et sélection**
  - Enrichissement lexicométrique
  - Évolution diachronique
  - Répartition entre types de corpus

sandrine.ollinger@atilf.fr