

---

■ ■ ■ ■ ■ ■ ■

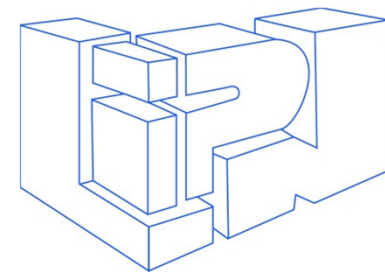
# Des traitements aux ressources linguistiques : le rôle d'une architecture linguistique

Frederik Cailliau

Villetaneuse, le 9 décembre 2010

Sous la direction d'Adeline Nazarenko

SINEQUA



# 1. Contexte

2. Problématique

3. Ressources linguistiques

4. Architecture linguistique

5. Environnement de gestion

6. Conclusion et perspectives

**Contexte**

Problématique

Ressources  
linguistiques

Architecture  
linguistique

Environnement  
de gestion

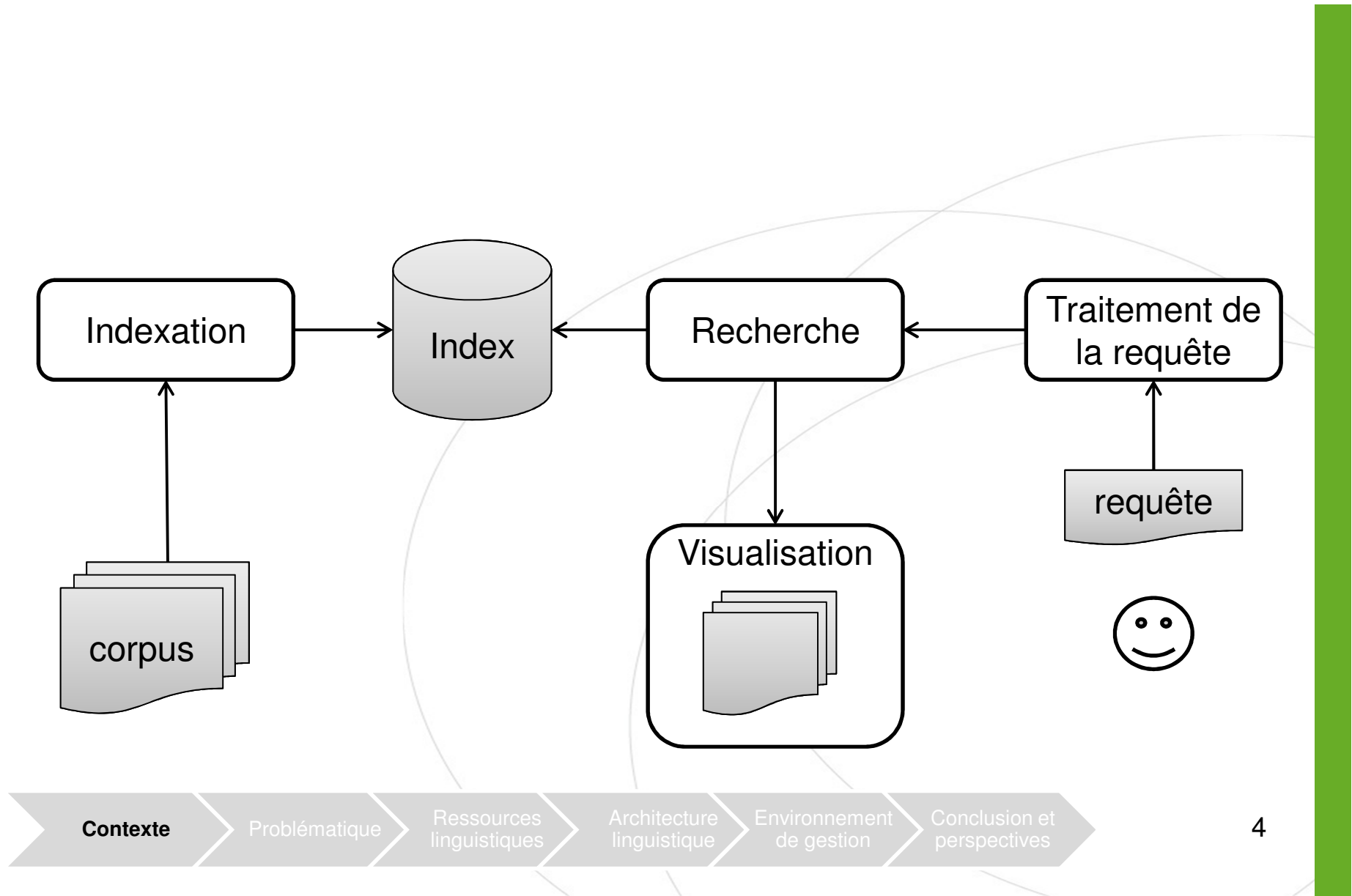
Conclusion et  
perspectives

# Sinequa

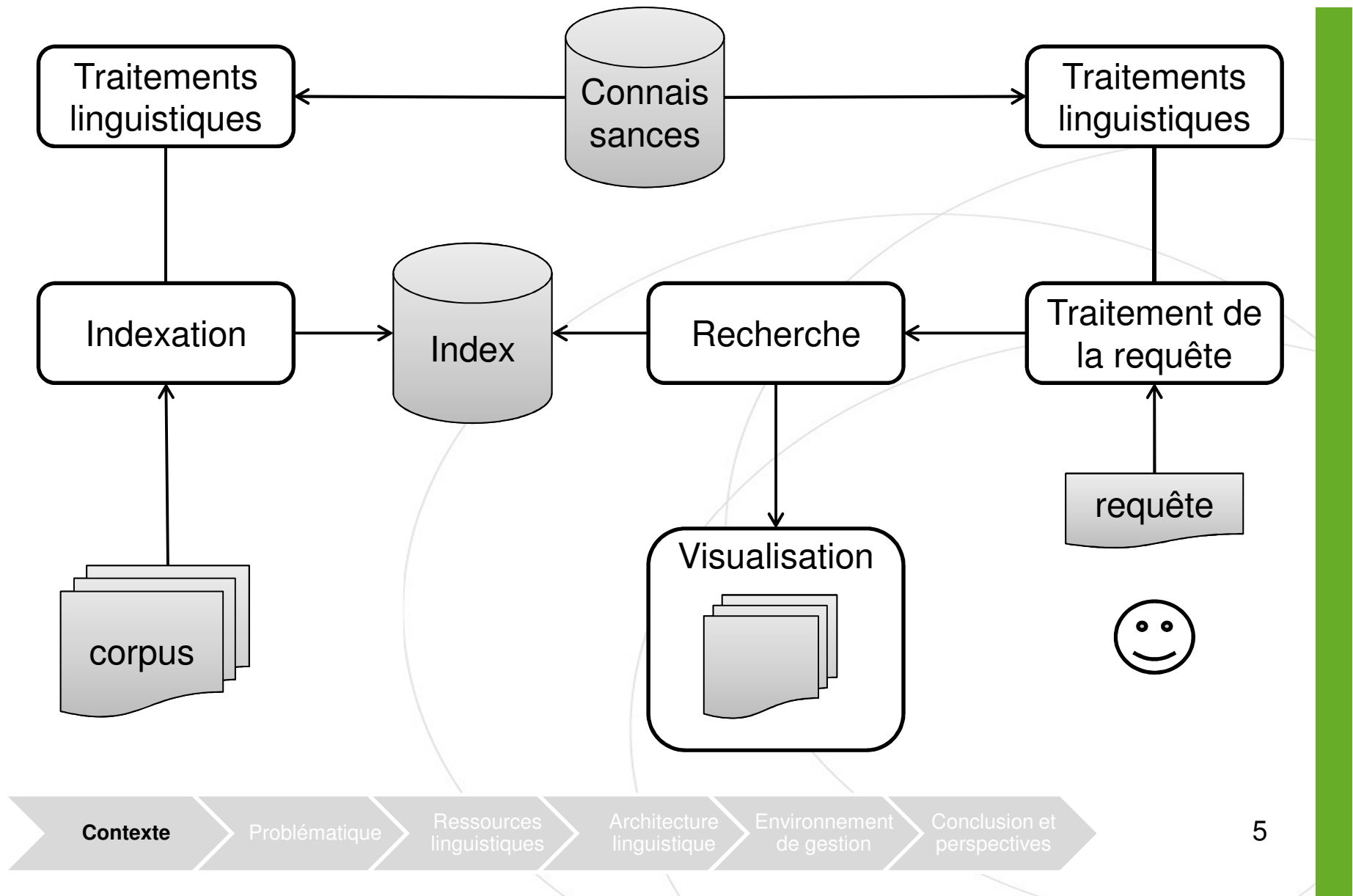
- Cifre à Sinequa
- Recherche d'Information  
moteur de recherche



# Moteur de recherche



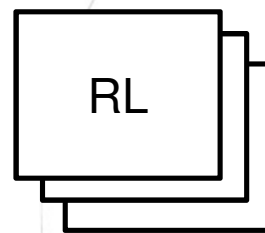
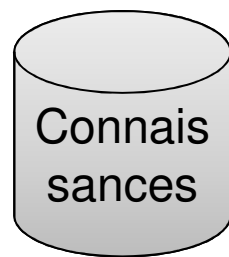
# Moteur de recherche



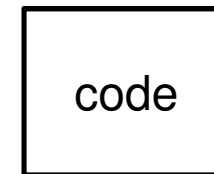


## Plate-forme d'annotation linguistique intégrée dans le moteur de recherche

- Code (générique) vs ressources (spécifique)



linguistes



informaticiens



# Contexte

## Contraintes industrielles

- Continuité du logiciel
- Non régression
- Rapidité des traitements
- Robustesse
- Cadre unique de gestion pour toutes les langues
- Généricité par rapport aux contextes applicatifs

Contexte


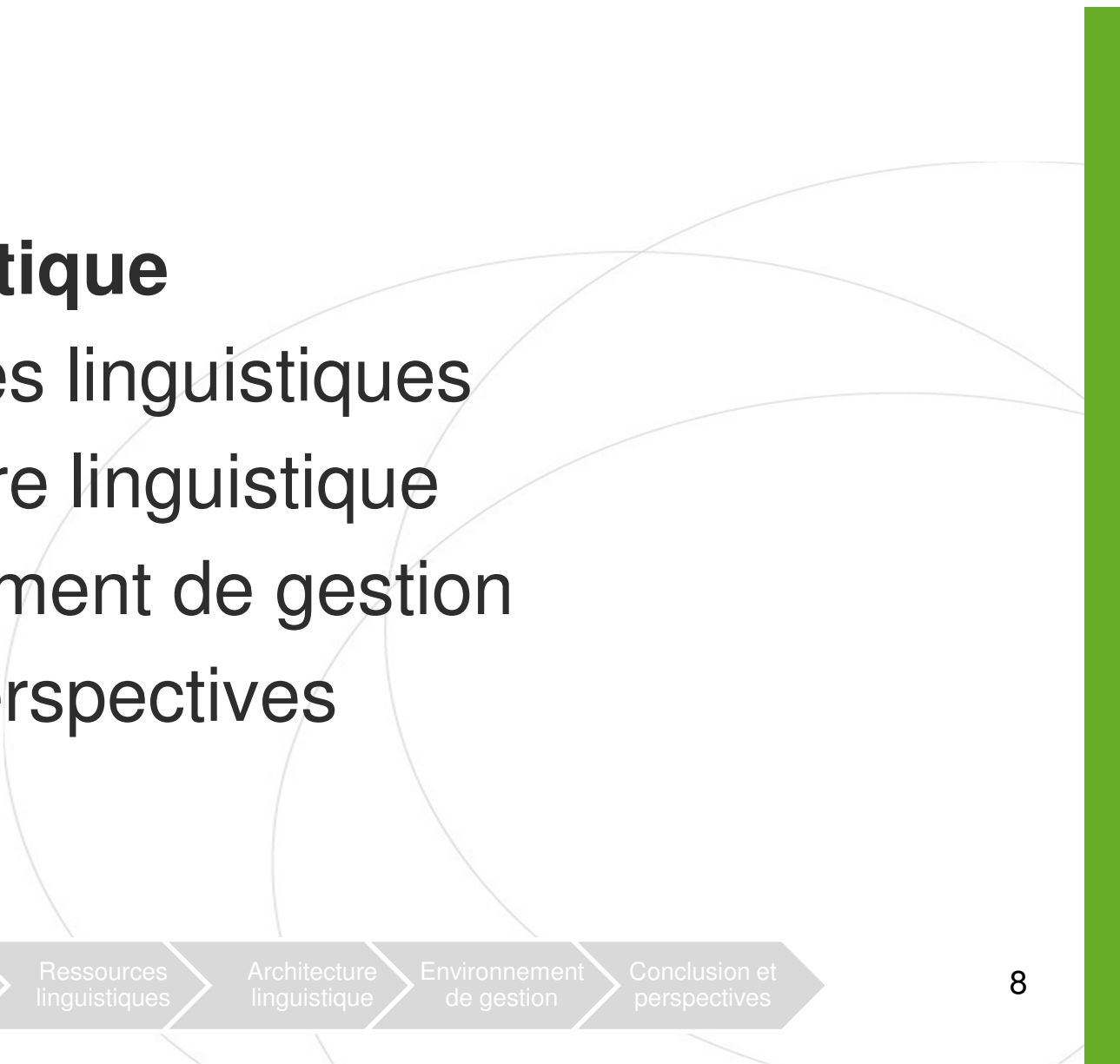
Problématique

Ressources  
linguistiques

Architecture  
linguistique

Environnement  
de gestion

Conclusion et  
perspectives

- 
- 
1. Contexte
  - 2. Problématique**
  3. Ressources linguistiques
  4. Architecture linguistique
  5. Environnement de gestion
  6. Bilan et perspectives
- 

Contexte

**Problématique**

Ressources  
linguistiques

Architecture  
linguistique

Environnement  
de gestion

Conclusion et  
perspectives



# Défi 1 : Volume

## Ressources

- Lexiques : > 1 500 fichiers ; 1,3 Go (texte)
- Grammaires : 128 automates, ...
- Corpus : 3 millions de fichiers ; > 50 Go

## Multitude de formats



# Défi 1 : Volume



## Connaissances

> 60 millions

- Lemmes (76 500)
- Mots-formes (x10)
- Descriptions (x1,5)
- Etiquettes (x3)
- Langues (x19)



# Défi 2 : Hétérogénéité des langues



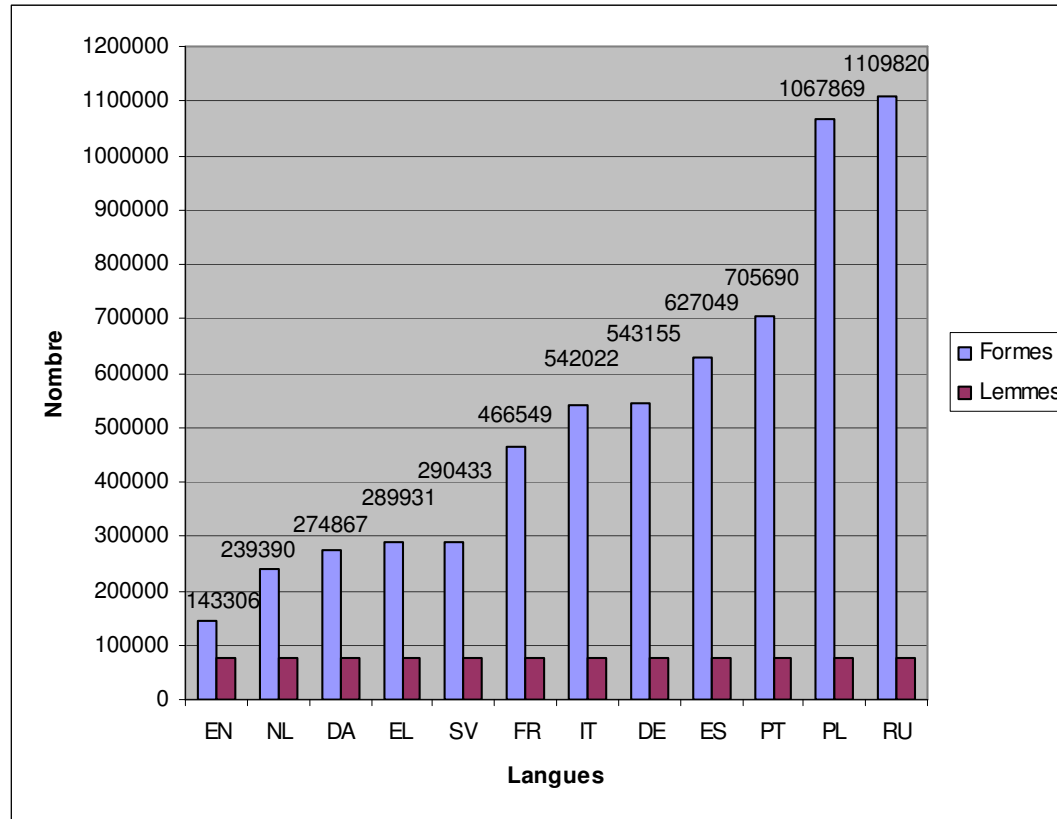
## 19 langues

- arabe, danois, allemand, grec, anglais, espagnol, finnois, français, italien, japonais, coréen, néerlandais, polonais, portugais, russe, suédois, thaï, chinois

→ Architecture générique



# Défi 2 : Hétérogénéité des langues



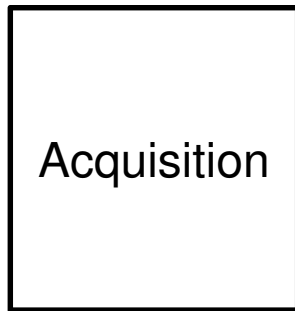
Projection du nombre de mots-formes pour un même nombre de lemmes

# Défi 3 : Evolution

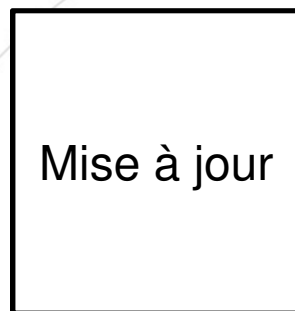


## Matière vivante

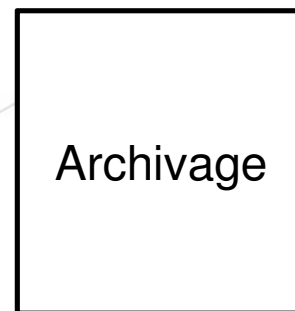
naissance



vie



mort



# Défi 4 : Cohérence

## Assurer la cohérence des connaissances

- intra-lexicale
- inter-lexicale
- entre lexiques et grammaires
- entre lexiques et corpus étiquetés



# Problématique

Comment gérer l'ensemble des ressources linguistiques pour une plate-forme d'annotation linguistique ?




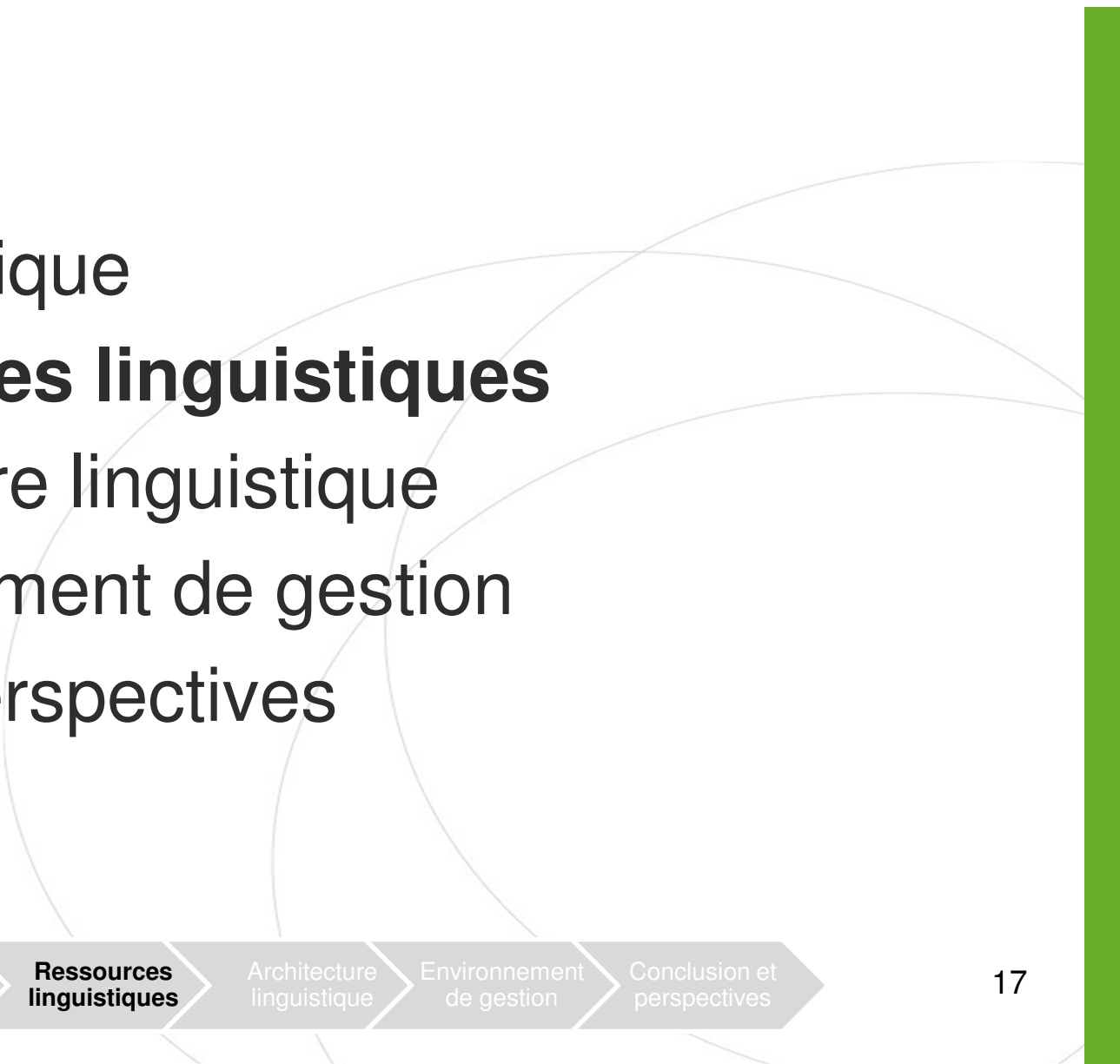
# Problématique

Comment gérer l'ensemble des ressources linguistiques pour une plate-forme d'annotation linguistique ?

- Architecture linguistique
- Environnement de gestion





- 
- 
1. Contexte
  2. Problématique
  - 3. Ressources linguistiques**
  4. Architecture linguistique
  5. Environnement de gestion
  6. Bilan et perspectives
- 

Contexte

Problématique

**Ressources  
linguistiques**

Architecture  
linguistique

Environnement  
de gestion

Conclusion et  
perspectives

# Ressource linguistique

Un ensemble de **données** comportant des **connaissances linguistiques** exploitables par un **traitement** automatique en particulier

- Lexiques
- Grammaires
- Corpus



# Lexiques

## Listes de mots avec des informations sur ces mots

manges : verbe singulier 2<sup>e</sup> personne, manger

est : verbe singulier 3<sup>e</sup> personne, être

est : nom masculin singulier, est

le : déterminant article défini singulier, le

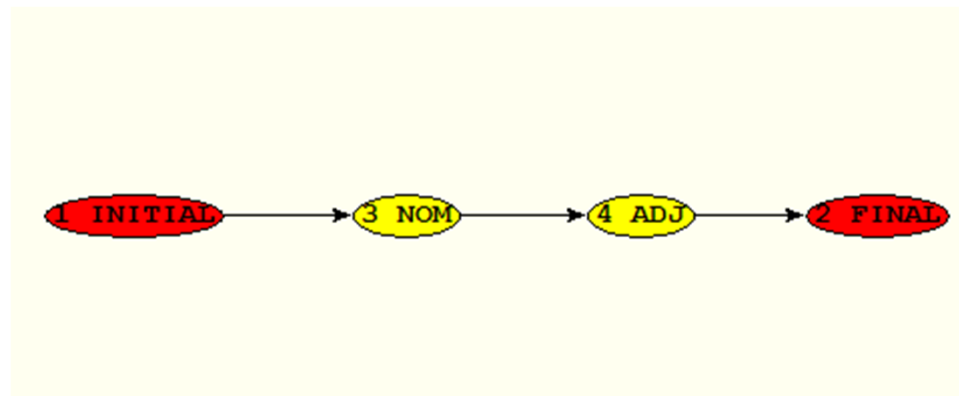
le : pronom singulier, le

# Grammaires

## Ensembles de règles

- Découpage en mots
- Flexion
- Décomposition de mots composés
- Analyse des mots dérivés
- Extraction d'entités (automates)

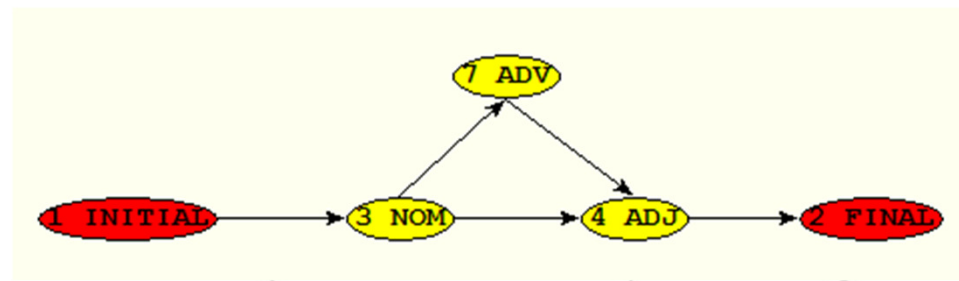
# Grammaires : automates



nom + adjectif

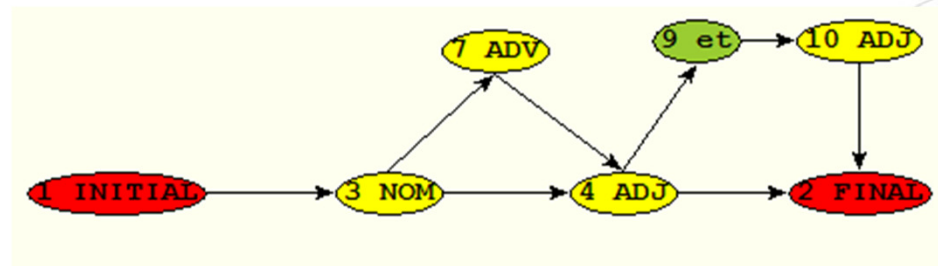
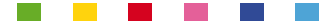


# Grammaires : automates



nom + adjectif  
nom + adverbe + adjectif

# Grammaires : automates



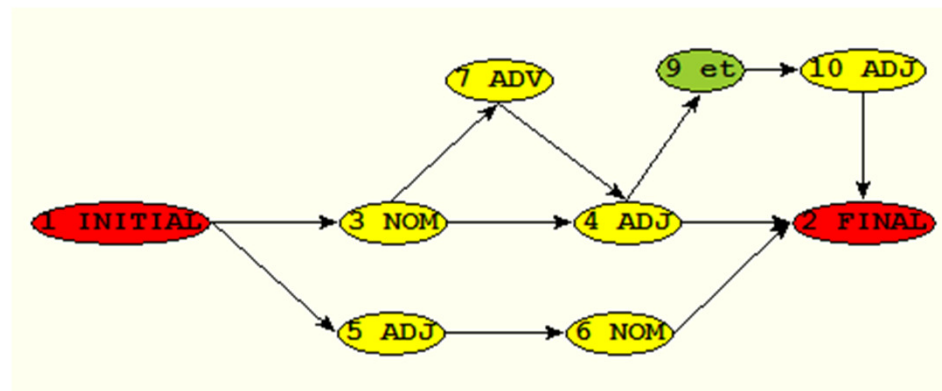
nom + adjectif

nom + adverbe + adjectif

nom + adjectif + et + adjectif

nom + adverbe + adjectif + et + adjectif

# Grammaires : automates



nom + adjectif

nom + adverbe + adjectif

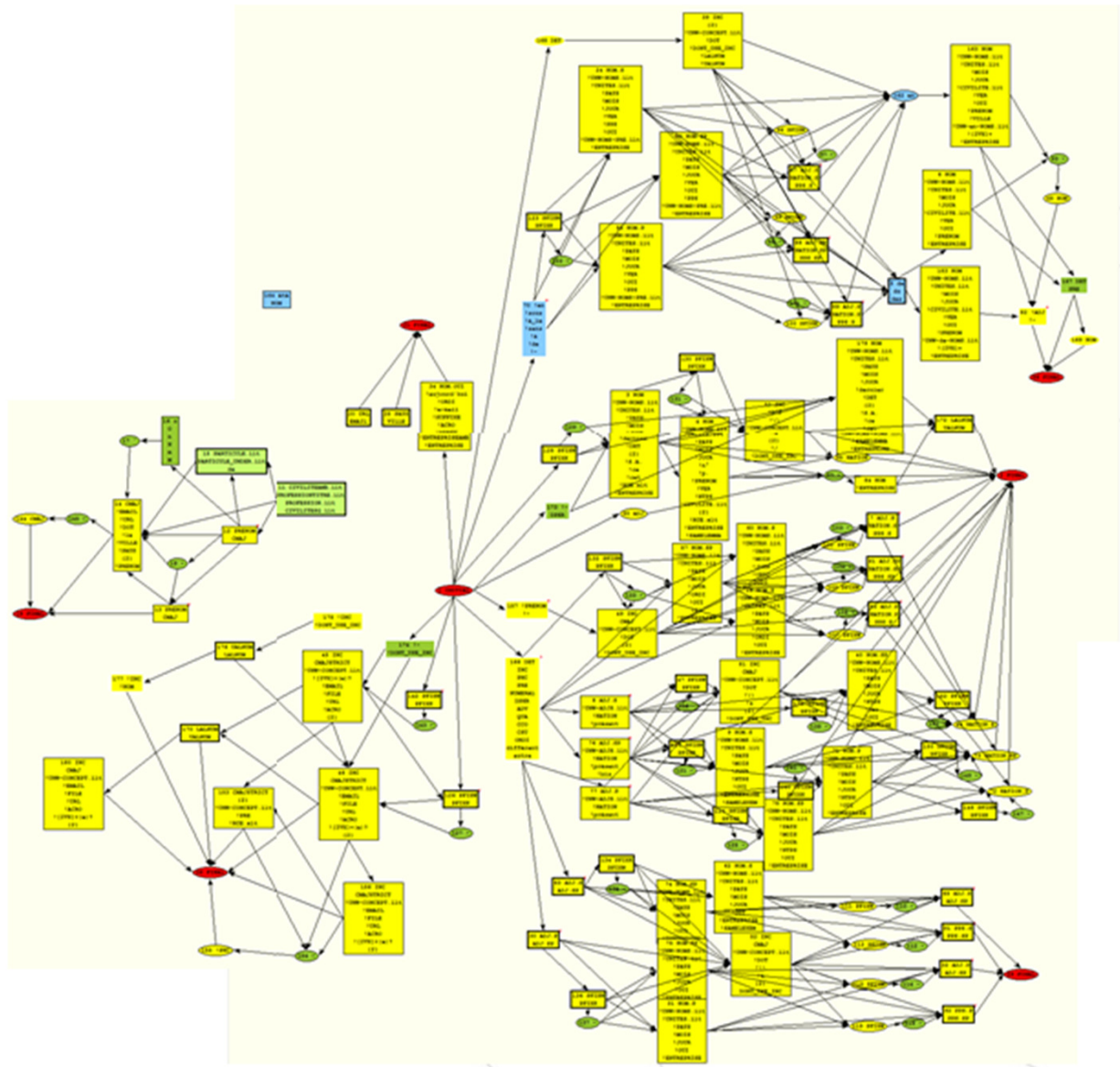
nom + adjectif + et + adjectif

nom + adverbe + adjectif + et + adjectif

adjectif + nom



# Grammaires : automates



# Corpus

Ensemble de documents réunis selon un critère spécifique

- Annoté : contenant des méta-informations

Il/pronom est/verbe beau/adjectif ./ponctuation

Contexte


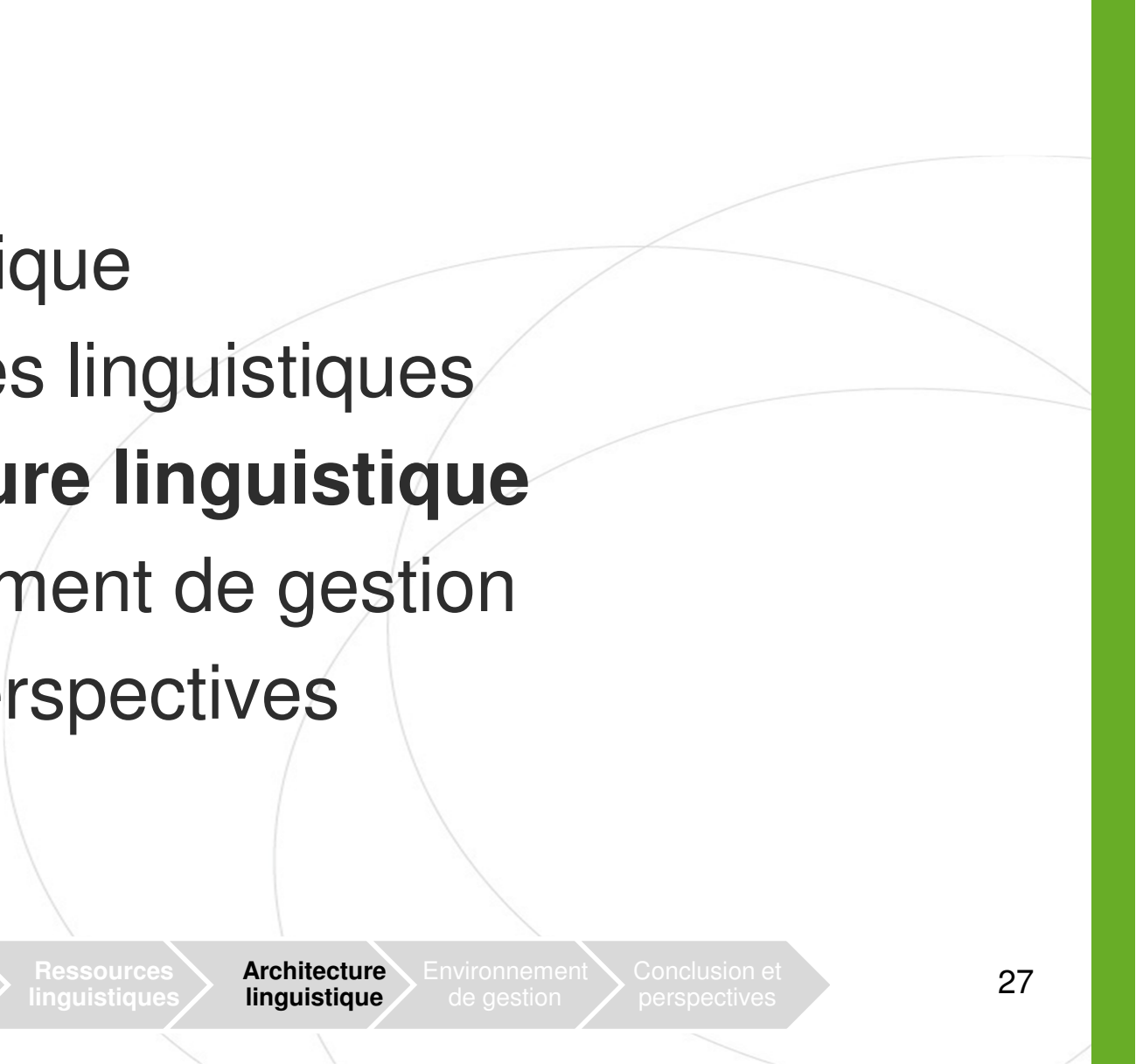
Problématique

**Ressources  
linguistiques**

Architecture  
linguistique

Environnement  
de gestion

Conclusion et  
perspectives

- 
- 
1. Contexte
  2. Problématique
  3. Ressources linguistiques
  - 4. Architecture linguistique**
  5. Environnement de gestion
  6. Bilan et perspectives
- 

Contexte

Problématique

Ressources  
linguistiques

**Architecture  
linguistique**

Environnement  
de gestion

Conclusion et  
perspectives

# Architecture linguistique

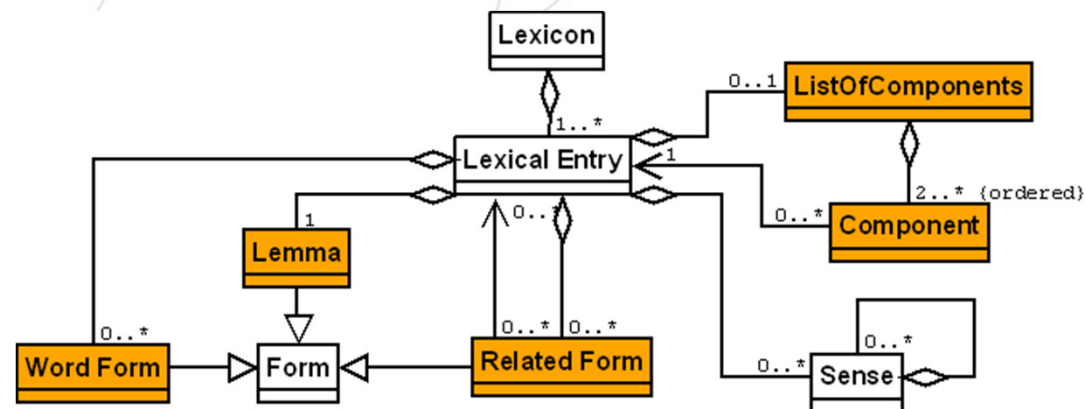
Sérasset (1994) : L'architecture linguistique définit les objets de base d'un dictionnaire et leurs relations.

Eagles (1993) : L'architecture linguistique définit les objets élémentaires du modèle et leurs relations. Elle spécifie aussi la terminologie générale commune au standard complet et utilisée pour discuter des dictionnaires, de leurs composants ou de l'interaction entre eux.

# Architecture linguistique

## Une longue histoire de travaux lexicaux

- Dictionnaires électroniques
- Projets de recherche (Genelex, Celex, etc)
- Norme ISO 24613:2008 (LMF)
  - DCR



# Architecture linguistique

Un modèle qui représente un ensemble de connaissances linguistiques ainsi que les relations entre ces connaissances.

Systeme :

- Traitements
- Ressources linguistiques : lexiques, grammaires, corpus



# Méthode

Recenser les types de connaissances utilisées dans le système et les mettre en relation

- Modèle ensembliste
- Visualisation : UML



# Traitements linguistiques

## Identifier les besoins en connaissances des traitements

- Identification de la langue
- Découpage en unités textuelles
- Analyse et étiquetage morphosyntaxique
- Désambiguïisation morphosyntaxique
- Lemmatisation
- Renvois entre unités textuelles
- Etiquetage sémantique
- Extraction d'entités

Contexte

Problématique

Ressources  
linguistiques

Architecture  
linguistique

Environnement  
de gestion

Conclusion et  
perspectives



# Architecture linguistique

- Décomposition

Rekrutierungsstelle = Rekrutierung + s + Stelle

- Règle

nom singulier + élément + nom = mot composé

- Connaissances

catégorie (unité lexicale)

traits morphologiques (unité lexicale)

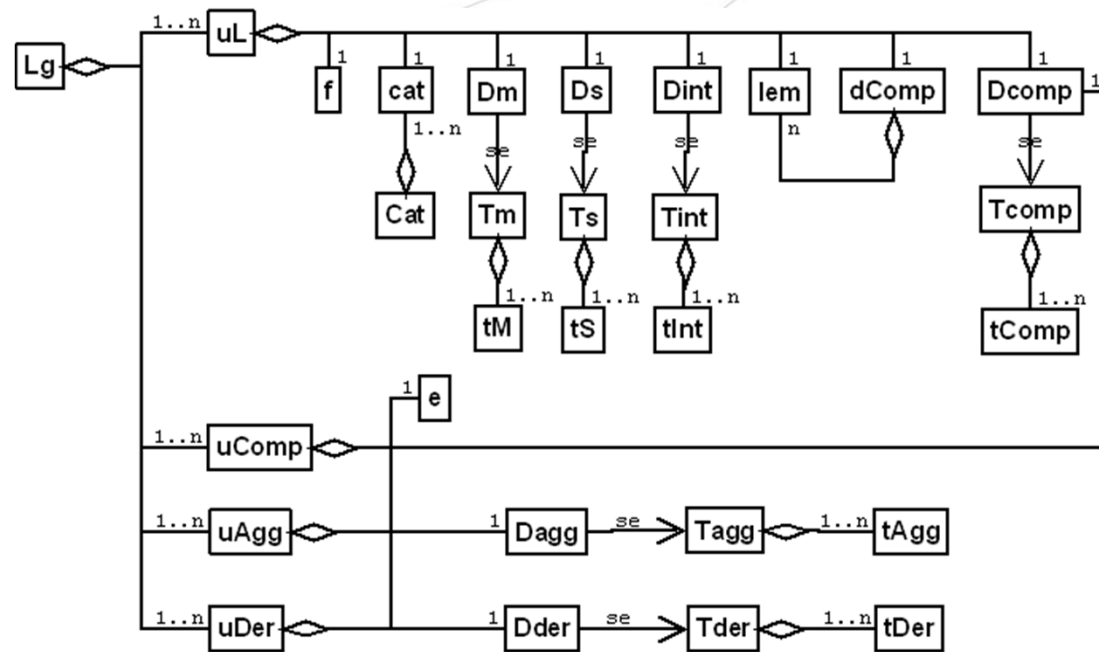
élément compositionnel (unité de composition)


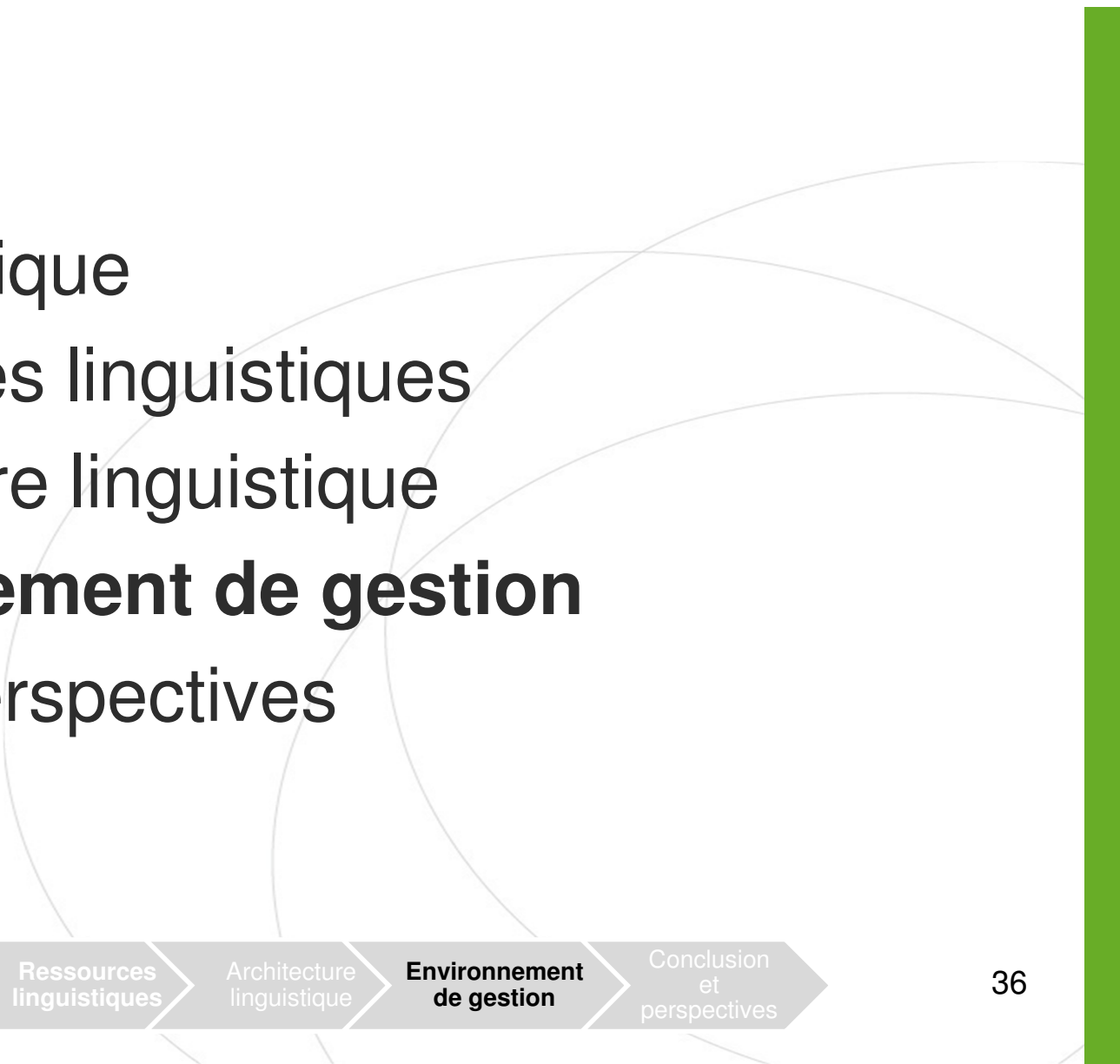
→ unité lexicale



# Architecture linguistique

- Modèle du *Lexique morphosyntaxique général*



- 
- 
1. Contexte
  2. Problématique
  3. Ressources linguistiques
  4. Architecture linguistique
  - 5. Environnement de gestion**
  6. Bilan et perspectives
- 

Contexte

Problématique

Ressources  
linguistiques

Architecture  
linguistique

**Environnement  
de gestion**

Conclusion  
et  
perspectives

# Environnement de gestion

## Composants

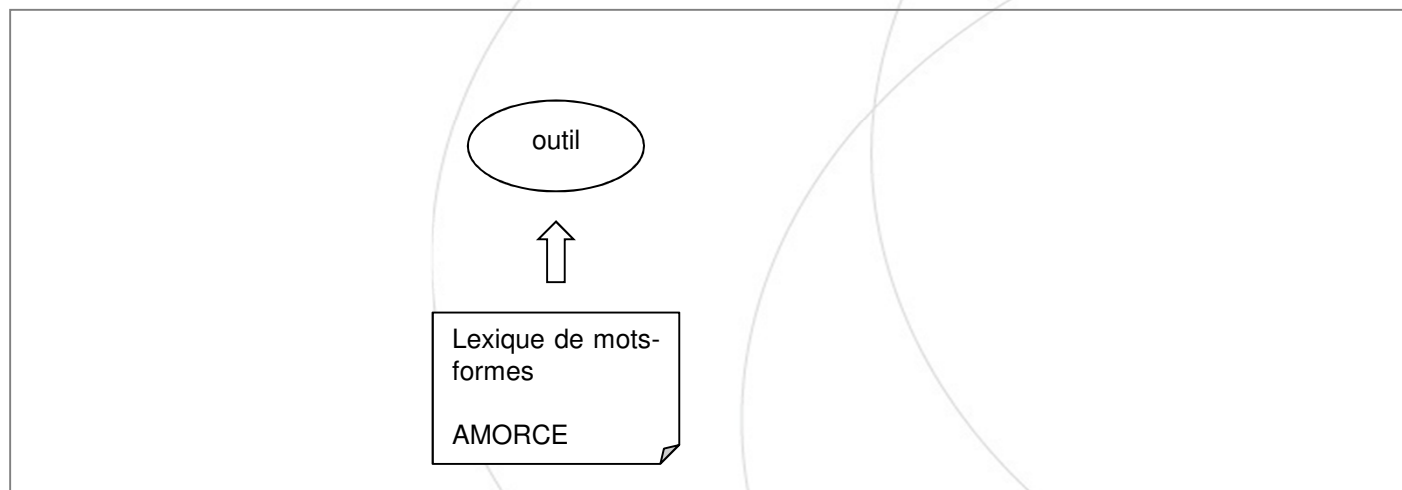
- Premiers composants
  - Système de *versioning* (SVN)
  - Documentation linguistique collaborative (81pages)
  - Scripts
  - Centralisation des corpus
- Outil d'acquisition
- Outil de mise à jour
- Outils de suivi



# Aide à l'enrichissement lexical

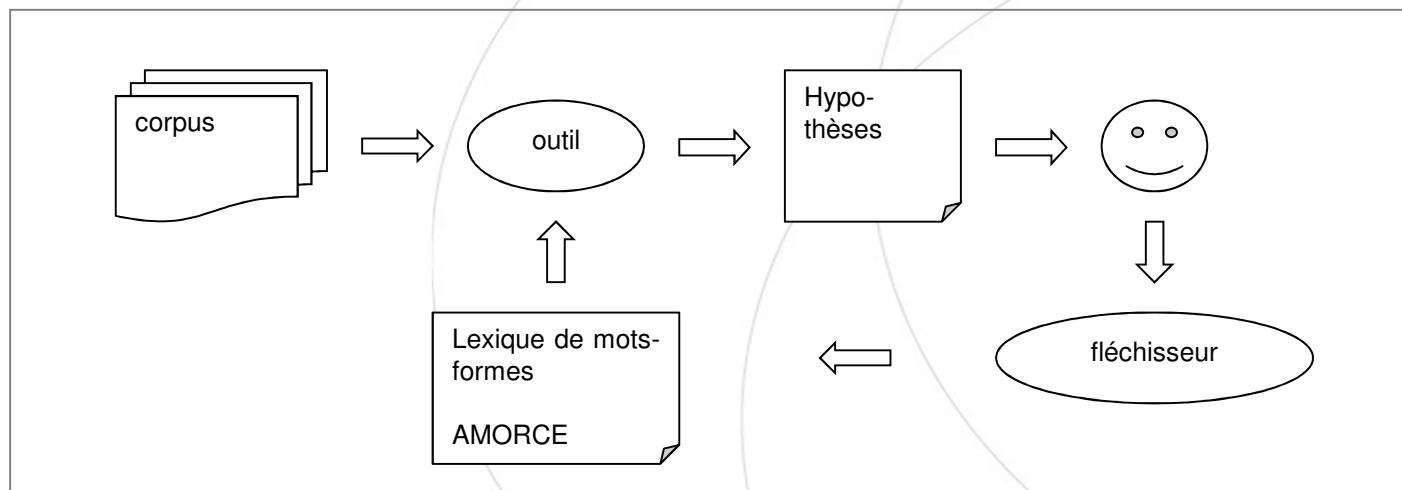


Proposer lemme et catégorie grammaticale à partir de la terminaison du mot pour les mots hors lexique



# Aide à l'enrichissement lexical

Proposer lemme et catégorie grammaticale à partir de la terminaison du mot pour les mots hors lexique



# Aide à l'enrichissement lexical



## 1. Construction de la liste des terminaisons avec fréquences à partir d'un lexique amorce

Terminaison-description	Fréquence
n-NOM.D.F.P	13731
n-NOM.N.F.P	13609
n-NOM.A.F.P	13601
n-NOM.G.F.P	13600
en-NOM.D.F.P	13447
en-NOM.N.F.P	13327
en-NOM.A.F.P	13319
en-NOM.G.F.P	13318
s-NOM.G.M.S	11174
t-VER.SUP.TU	9612
st-VER.SUP.TU	9612
e-VER.SUP.IL	9598
e-VER.SUP.JE	9590
...	...

1 764 832 terminaisons  
870 096 term. uniques

...		
mokratisieren-VER.SUP.ILS		1
etuckerten-VER.PPS.EN		1
chäftiger-ADJ.ER		1
ßtischstes-ADJ.SUPER.ES		1
msiedlungen-NOM.G.F.P		1
ammenschlüsse-NOM.A.M.P		1
hndung-NOM.G.F.S		1
tent-NOM.D.X.S		1





# Aide à l'enrichissement lexical

1. Construire la liste des terminaisons avec fréquences à partir du lexique
- 2. Filtrage : seuil de fréquence (10)**

= élimination des terminaisons longues

1 764 832 → 47 854 terminaisons

870 096 → 22 081 terminaisons uniques

# Aide à l'enrichissement lexical

1. Construire la liste des terminaisons avec fréquences à partir du lexique
2. Application d'un seuil de fréquence
3. **Réduction du nombre de terminaisons selon le score d'entropie**

telnder, ppelnder, tzelnder, ckelnder, ndelnder, ügelnder, sselnder, ickelnder, nzelnder, mmelnder, Inder  
→ Inder

22 081 → 6 962 terminaisons uniques

# Aide à l'enrichissement lexical

1. Construire la liste des terminaisons avec fréquences à partir du lexique
2. Application d'un seuil de fréquence
3. Calcul du pouvoir informationnel de chaque terminaison et réduction du nombre de terminaisons ayant le même score d'entropie
4. **Construction des hypothèses sur le corpus**

# Aide à l'enrichissement lexical



Fréq. racine	Racine	Terminaison
2	Rekrutierungsst	elle
2	Rekrutierungsst	ellen
2	Allosaur	ier
2	Allosaur	us
2	Agrarwirts	chaft
2	Agrarwirts	chaften
2	Ahmadinescha	d
2	Ahmadinescha	ds

Rekrutierungsst      elle

[ADJ.E-81][NOM.A.F.S-64][NOM.A.M.P-14]  
 [NOM.A.X.P-19][NOM.D.F.S-59][NOM.G.F.S-59]  
 [NOM.G.X.P-19][NOM.N.F.S-64][NOM.N.M.P-14]  
 [NOM.N.X.P-19][VER.IM.TU-95][VER.PI.JE-97]  
 [VER.SUP.IL-97][VER.SUP.JE-97]

Rekrutierungsst      ellen

[ADJ.EN-81][NOM.A.F.P-68][NOM.D.F.P-68]  
 [NOM.D.M.P-16][NOM.D.X.P-21][NOM.G.F.P-68]  
 [NOM.N.F.P-68][VER.IN-97][VER.PI.ILS-97]  
 [VER.PI.NOUS-97][VER.SUP.ILS-97]  
 [VER.SUP.NOUS-97][VER.ZU-74]

# Aide à l'enrichissement lexical



Fréq. racine	Racine	Terminaison
2	Rekrutierungsst	elle
2	Rekrutierungsst	ellen
2	Allosaur	ier
2	Allosaur	us
2	Agrarwirts	chaft
2	Agrarwirts	chaften
2	Ahmadinescha	d
2	Ahmadinescha	ds

Rekrutierungsst      elle

[ADJ.E-81][NOM.A.F.S-64][NOM.A.M.P-14]  
 [NOM.A.X.P-19][NOM.D.F.S-59][NOM.G.F.S-59]  
 [NOM.G.X.P-19][NOM.N.F.S-64][NOM.N.M.P-14]  
 [NOM.N.X.P-19][VER.IM.TU-95][VER.PI.JE-97]  
 [VER.SUP.IL-97][VER.SUP.JE-97]

Rekrutierungsst      ellen

[ADJ.EN-81][NOM.A.F.P-68][NOM.D.F.P-68]  
 [NOM.D.M.P-16][NOM.D.X.P-21][NOM.G.F.P-68]  
 [NOM.N.F.P-68][VER.IN-97][VER.PI.ILS-97]  
 [VER.PI.NOUS-97][VER.SUP.ILS-97]  
 [VER.SUP.NOUS-97][VER.ZU-74]

# Evaluation

Précision = mots lexicalement intéressants  
mots proposés

- Mots non capitalisés
  - entre 70% et 85% pour les mots dont le radical apparaît 3 à 7 fois
  - entre 55% et 70% pour les mots dont le radical apparaît 7 à 11 fois
- Mots capitalisés
  - environ 10%



# Environnement de gestion : acquisition

## Résumé

- Méthode originale utilisant les informations à disposition
- Ajout de lexique + découverte erreurs lexique
- Seuils paramétrables : influent sur la précision et le rappel



# Environnement de gestion : acquisition

## Bilan

- Bons résultats pour les verbes en allemand
- La capitalisation des noms communs en allemand perturbe les résultats pour les noms
- A tester sur d'autres langues (flexionnelles)
- Ajout d'autres paramètres, ex. fréquence des mots dans le corpus





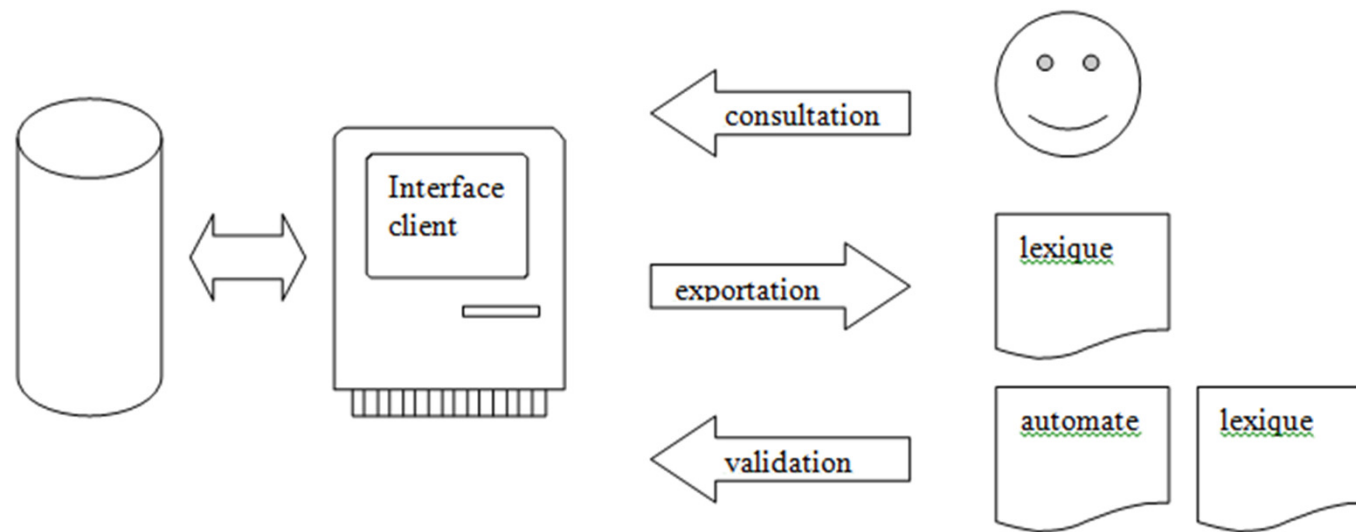
# Mise à jour

## Modification de ressource

- Demande client
- Signalement de bug
- Changement de politique



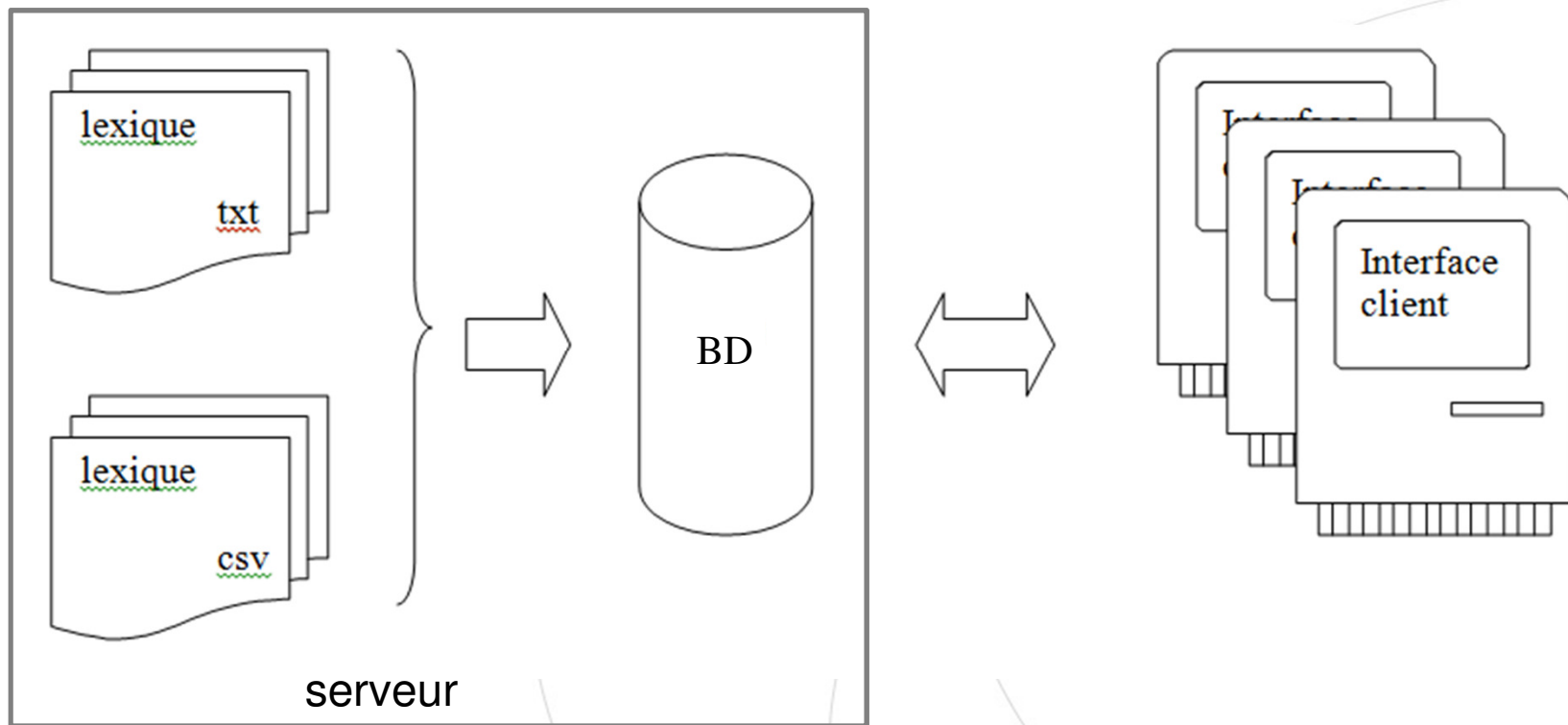
# Outil d'édition lexicale



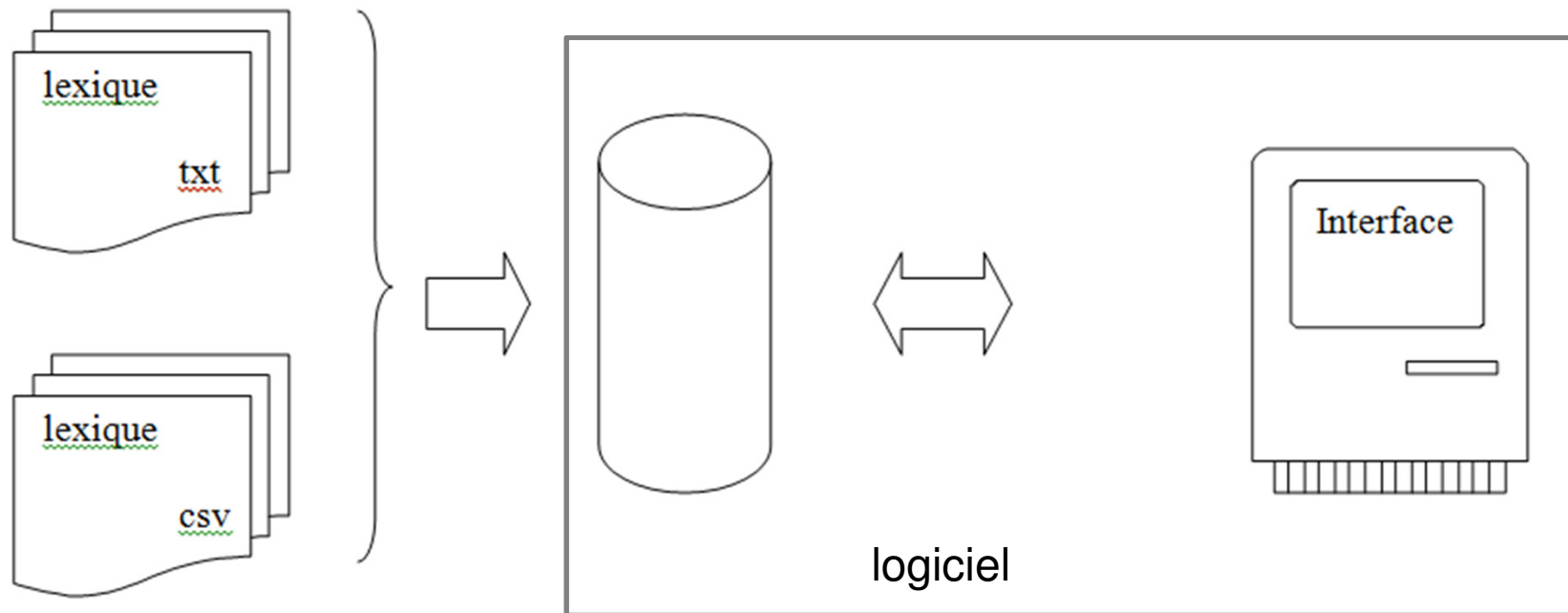
- Tests de cohérence intra- et inter-lexicale
- Documentation intégrée



# Architecture du prototype



# Industrialisation du prototype



# Modification de l'architecture logicielle

- Suivi de modifications et des versions
- Multi-utilisateur peu important à Sinequa
- Evolution puissance de calcul de machines utilisateurs



# Outil d'édition lexicale : interface



	Dst File	Fom	Lemma	Semantic	Label	Comment
<input type="checkbox"/>	src/dst/base_en...	hindu	hindu	446	ADJ	
<input type="checkbox"/>	src/dst/base_en...	hindu	hindu	446	NOM.S	
<input type="checkbox"/>	src/dst/base_en...	hinduism	hinduism	446	NOM.S	
<input type="checkbox"/>	src/dst/base_en...	hinduisms	hinduism	446	NOM.P	
<input type="checkbox"/>	src/dst/base_en...	hinduize	hinduize		VER.BASE	
<input type="checkbox"/>	src/dst/base_en...	hinduized	hinduize		PPS	
<input type="checkbox"/>	src/dst/base_en...	hinduized	hinduize		VER.PAST	

234546 row(s) selected

**Logs**

- Loading dst src/dst/base\_en.dst...
- Loading dst src/dst/geo\_monde.en.dst...
- Loading dst src/dst/prenoms.en.dst...

Dictionary en - english loaded

# Outil d'édition lexicale

- Bilan

- Outil en place et maintenu
- Gain de temps
  - Recherche rapide
  - Recherches complexes (scripts)
- Communication
  - Qualification d'erreur en liaison avec le service support



# Pilotage

- Outils destinés au superviseur
  - Capitaliser l'expérience
- 2 outils
  - Calcul de la complexité des automates
  - Suivi de l'évolution des lexiques





# 1<sup>er</sup> outil de pilotage

## Calcul de la complexité des automates

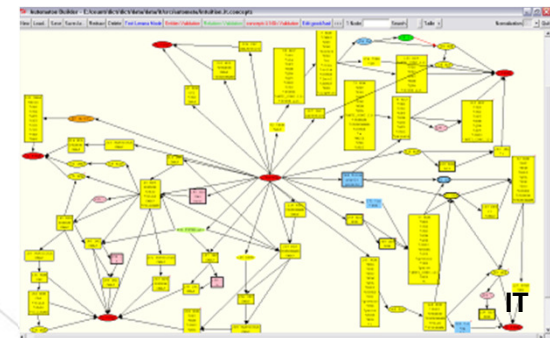
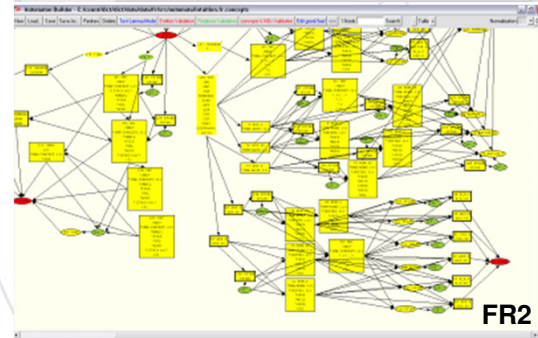
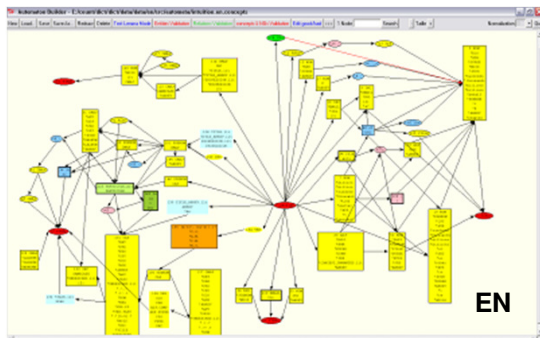
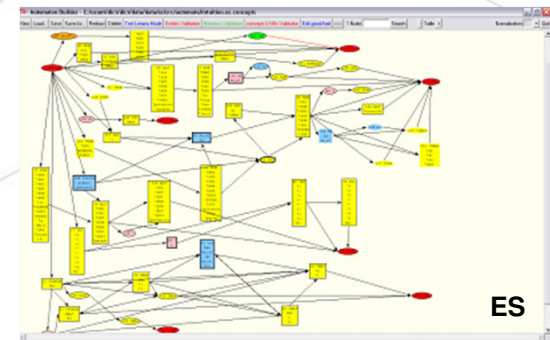
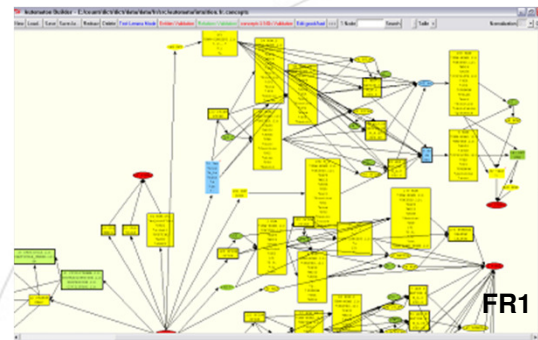
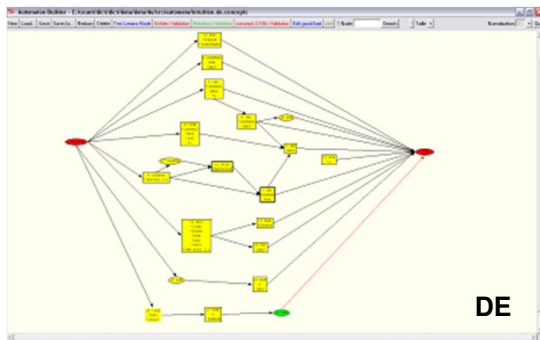
- Un automate est simple, donc plus facile à gérer, si le nombre de chemins est proche ou inférieur au nombre de transitions. (linéarité des chemins)
- Un automate est simple si le nombre de transitions est proche du nombre de nœuds.

$$\begin{aligned} \text{Complexité}(Aut) &= \sqrt{\frac{\#chemins(Aut)}{\#transitions(Aut)} \times \frac{\#transitions(Aut)}{\#noeuds(Aut)}} \\ &= \sqrt{\frac{\#chemins(Aut)}{\#noeuds(Aut)}} \end{aligned}$$

# 1<sup>er</sup> outil de pilotage



	Nœuds	Transitions	Chemins	Complexité
DE	22	35	17	0,88
EN	61	112	884	3,81
ES	58	102	175	1,74
FR	154	379	1424	3,04
IT	74	127	285	1,96



# 2<sup>nd</sup> outil de pilotage

## Evolution des ressources

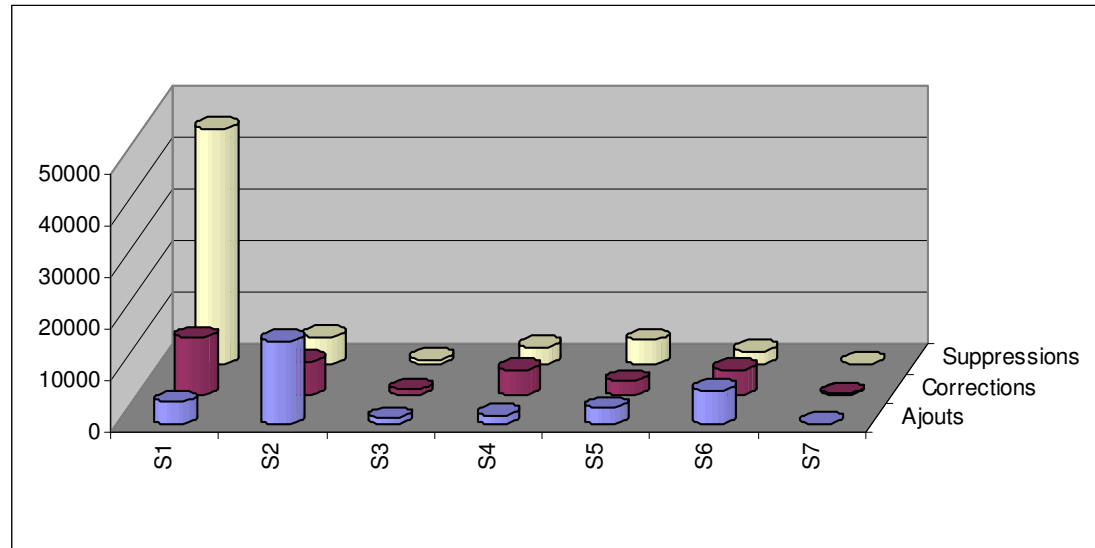
- Informations sur les modifications (SVN)
- Heuristiques de différenciation entre ajout, modification et suppression



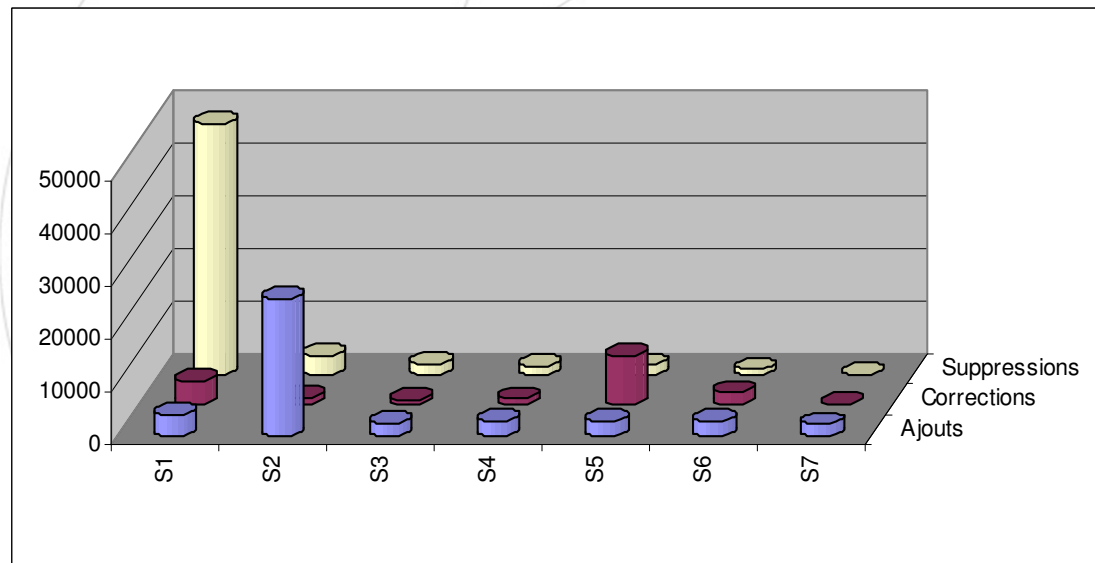
# Suivi de l'évolution des ressources



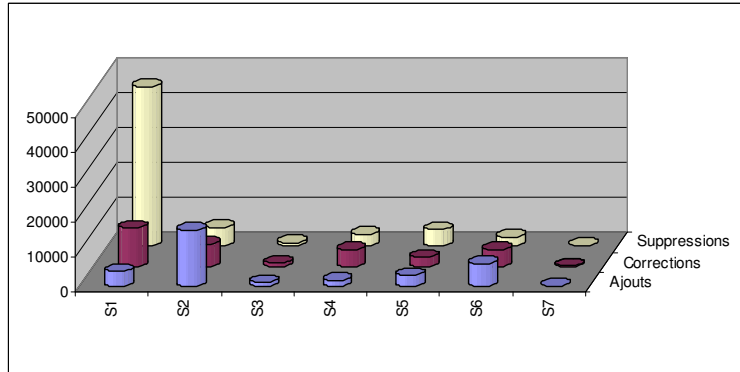
- FR



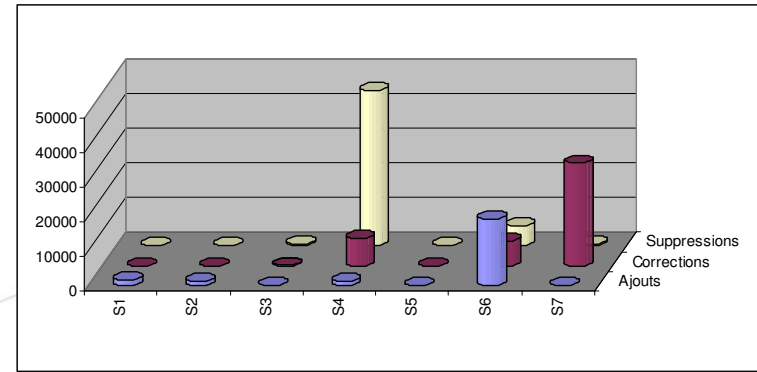
- EN



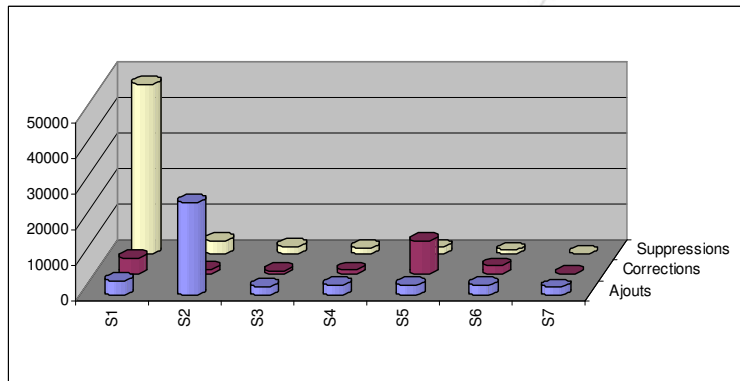
# Suivi de l'évolution des ressources



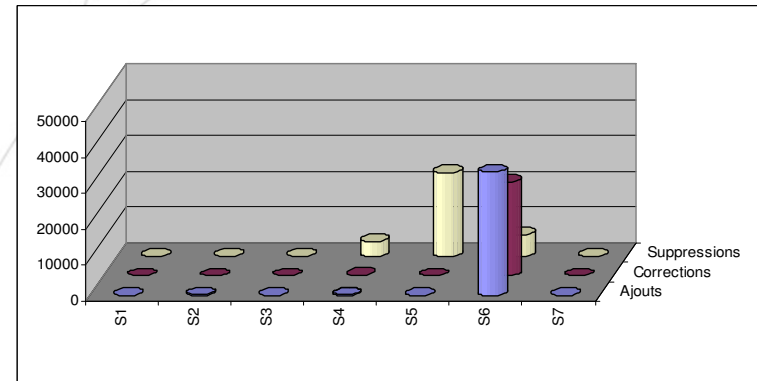
FR



ES



EN




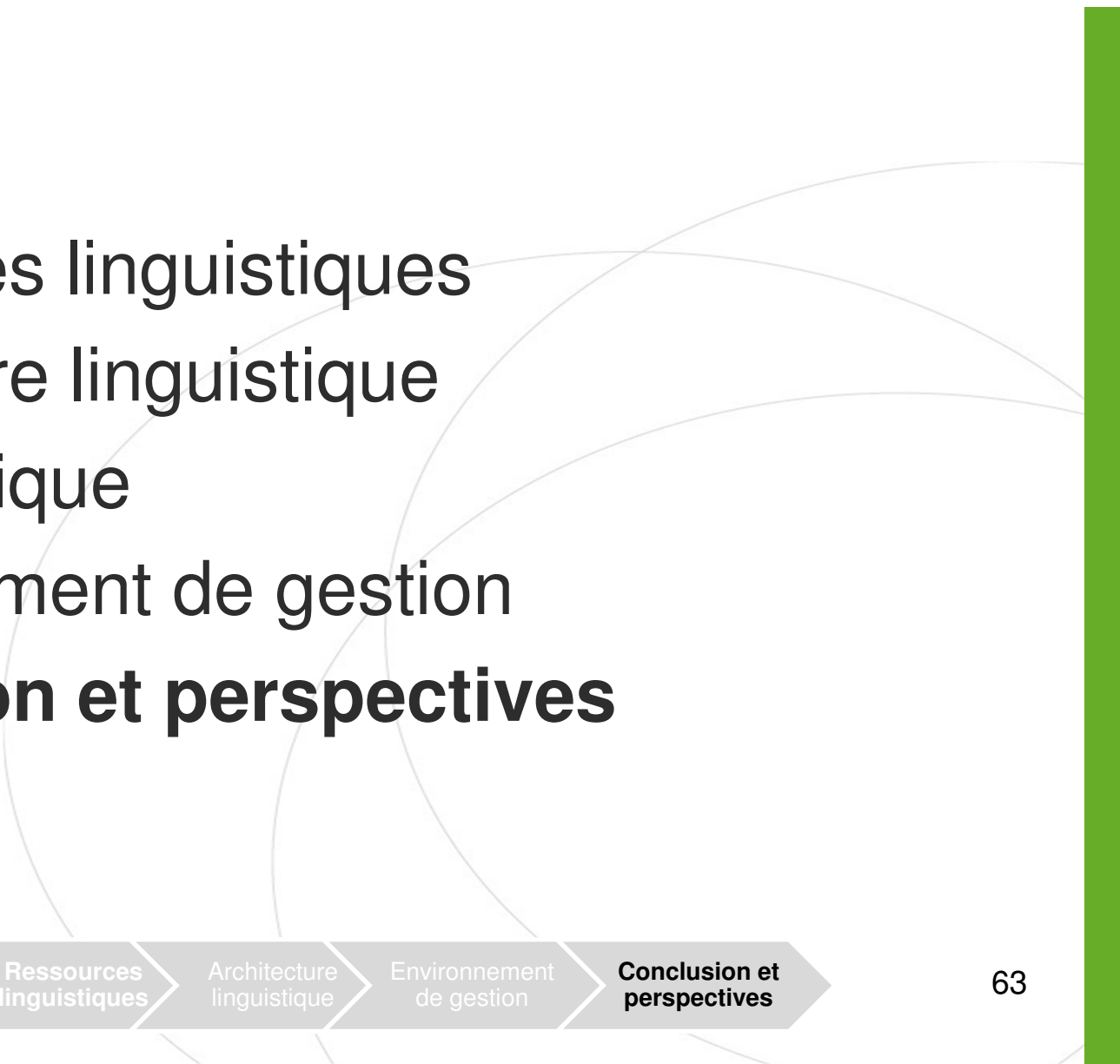
IT

# Outils de pilotage

- Bilan

- Evolution : un important effort d'interprétation est nécessaire.
- Si les composants de base sont en place, il est simple de mettre en place des outils de suivi global.



- 
- 
1. Contexte
  2. Ressources linguistiques
  3. Architecture linguistique
  4. Problématique
  5. Environnement de gestion
  - 6. Conclusion et perspectives**
- 

Contexte

Problématique

Ressources linguistiques

Architecture linguistique

Environnement de gestion

**Conclusion et perspectives**

# Conclusion

- Description de l'architecture linguistique du système
- Méthode générique pour expliciter l'architecture linguistique d'un système
  - Compréhension totale du système : interaction entre les traitements et les ressources
  - Partage du savoir
  - Partage des outils et des ressources





# Conclusion

- Environnement
  - Simplifiant la gestion au jour le jour
    - gain de temps
  - Augmentant la cohérence des données
    - limitant le nombre de retours clients dans la durée
  - Outils de pilotage
    - partage des savoirs

Contexte

Problématique

Ressources linguistiques

Architecture linguistique

Environnement de gestion

Conclusion et perspectives

# Perspectives

- Compléter l'environnement
  - Continuer l'industrialisation de l'outil d'édition : accès en écriture, validation des automates
  - Intégrer l'outil d'acquisition dans l'outil d'édition lexicale



# Perspectives

- Réduire le nombre de connaissances
  - Conversion des lexiques morphosyntaxiques en extension vers des lexiques en intension



# Perspectives

- Développer des mesures pour évaluer la dégradation des traitements linguistiques dans des contextes non standard (transcriptions automatiques...) en vue d'adapter les ressources linguistiques



---



Merci de votre attention.