Séminaire sur le multilinguisme pour l'accès aux contenus

Orange Labs Lannion

30 juin 2009

Organisateurs:

Malek Boualem TECH/ACTS/FAST

Benoît Gaillard TECH/ACTS/FAST

orange™

Sommaire

1. Introduction (9h00-9h45) :
   - 09h00-09h05 : Introduction au séminaire (M.Boualem).
   - 09h10-09h20 : Multilinguisme pour l'accès aux contenus (H.Sanson).
   - 09h25-09h40 : Traitement des langues et accès aux contenus (G.Prigent).

2. Travaux en Open Innovation sur le CLIR dans QUAERO-MSSE (9h45-10h30) :
   - 09h45-09h55 : Démonstration du prototype multilingue MSSE (M.Boualem & al.).
   - 10h00-10h10 : Sources d'information et connecteurs multilingues à VSE (P.Diverres).
   - 10h10-10h20 : Architecture du prototype multilingue MSSE (Y.Almeras).
   - 10h20-10h30 : Roadmap de mise en œuvre du CLIR dans MSSE (B.Gaillard, M.Boualem).

3. Evaluation, besoins et usages du multilinguisme (10h30-11h00) :
   - 10h30-10h40 : Evaluation de moteurs intégrant une dimension multilingue (B.Gaillard).
   - 10h45-10h55 : Besoins et usages du multilinguisme dans les projets d'Orange (M.Boualem).

4. Traitement des requêtes pour l'accès aux contenus multilingues (11h00-11h40) :
   - 11h00-11h10 : Pré-traitement et traduction des requêtes (B.Gaillard, M.Boualem).
   - 11h15-11h25 : Expansion des requêtes pour le multilinguisme (J.L.Bouraoui, E.Guimier de Neef).
   - 11h30-11h40 : Interprétation des requêtes à base d'ontologies (J.Heinecke, E.Guimier de Neef).

5. Multilinguisme pour 24/24 Actu (11h45-12h00) :
   - 11h45-11h55 : Expérimentation bilingue pour 24/24 Actu (P.Filoche et T.Urvoy).

Sommaire

1. Introduction (9h00-9h45) :

   – 09h10-09h20 : Multilinguisme pour l'accès aux contenus (H.Sanson).

Opportunités relative au CLIR vues de l'OR Media Search

- Transcription de contenus en VO sur sites UGC (études Content Lab 2007)
  - L'attente des internautes est de contourner la chronologie géographique de la distribution des films et séries => pas la position d'Orange.
  - Applications "légales" restent à identifier (la traduction étant un élément de la localisation du contenu et fait partie du cycle d'exploitation pour les contenus Premium
- L'accès aux sources multimédia d'information étrangères
  - Applications B2C: Application 24/24 Actu: à préciser Next.com
  - Applications B2B: intelligence économique, communication entreprise, …
- L'accès multilingue aux vidéos semi-pro:
  - Fonds culturels (trop cher à traduire à la main)
  - Vidéo touristiques, pratiques
  - Vidéo scientifique, et plus généralement destinées à diffuser la connaissance (médecine, finance, ….), de plus en plus présentes sur sites UGC
- Local search ?
  - Traduction requêtes et/ou contenus lors de l'accès en roaming à des ressources locales à l'étranger (business search, city guide, ….)
- Conclusion: besoin d'approfondir la recherche et l'analyse des opportunités

Sommaire

1. Introduction (9h00-9h45) :

   – 09h25-09h40 : Traitement des langues et accès aux contenus (G.Prigent).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion, 30 juin 2009

Traitement des langues et accès aux contenus

Gilles Prigent , TECH/ACTS/FAST

orange™

Traitement des langues et accès aux contenus

partie 1 actions de l'urd FAST

partie 2 illustration : étude de corpus de requêtes

actions de l'urd FAST

Traitement des langues et accès aux contenus l'urd TECH/ACTS/FAST

- Laboratoire ACTS : Acces to Content Technologies & Services
- Urd FAST : Future ArchitectureS & textual Technologies
- Mission principale :
  - Apporter une expertise technique et des solutions dans le domaine de traitement du texte et des architectures futures des services d'accès aux contenus
- Activités dans le domaine du traitement du texte pour des services innovants :
  - expertise technique, évaluations des solutions
  - développement d'outils technologiques
  - expertise pour l'intégration

Actions dans le domaine du traitement du texte
Deux grandes directions

- L'interaction : énoncés courts
  - requêtes :
    - typage, correction, normalisation, expansion, interprétation, recherche interlingue
  - SMS : normalisation avant vocalisation ou traduction

- Les contenus : contenus textuels ou métadonnées textuelles
  => enrichissement
  - normaliser : mots et expressions, unités de mesures, lemmatisation
  - extraire : mots-clés, entités nommées, termes, concepts, thèmes
  - découper : langue, thématique
  - classer : langue, thématique, sujet

☞ compétences : informatique, linguistique, statistique

☞ capitalisation
  - logiciels et backoffice (tilt)
  - ressources linguistiques, corpus, bases de données

Actions dans le domaine du traitement du texte
Contribution aux projets

- OR Media Search :
  - Projets pour le media, local et mobile search
  - Projets collaboratifs : Pharos, Quaero
- 3P Search & Advertizing
  - Local search, annuaire : 118712
  - Correction et interprétation de requêtes : OPF
  - NewsCloud
- 3P TV Services for Mobile
  - My TV VOD Adviser Mobile

illustration : étude de corpus de requêtes

illustration : étude de corpus de requêtes

- Objectif : mettre en évidence par l'analyse de corpus de requêtes de phénomènes nécessitant une analyse linguistique évoluée
- Étude basée sur une analyse de plusieurs corpus de requêtes
- Réalisée par
  - Benoît Gaillard, TECH/ACTS/FAST
  - Jean-Léon Bouraoui, TECH/ACTS/FAST
  - Emilie Guimier de Neef, TECH/ACTS/FAST
  - Malek Boualem, TECH/ACTS/FAST

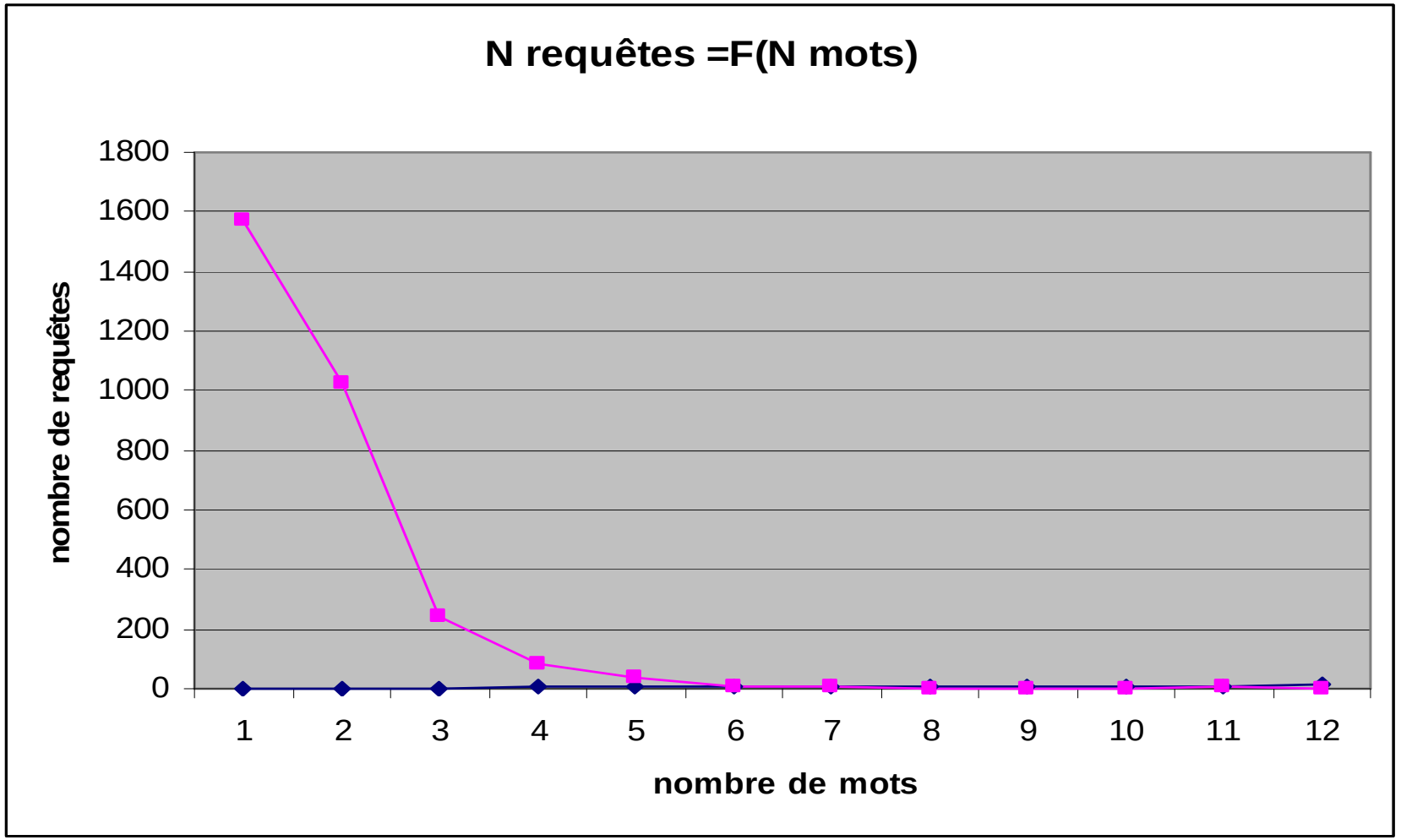
Corpus analysés

- VSE
  - 3380 requêtes:
  - peu de redondances, S2 2008. Vidéos, actualités. Panel interne.
- Ma Zone Video
  - Cible: vidéos, UGC, domaine général, public
  - 10 000 requêtes classées par fréquence.
- Top Mobile
  - Mobiles, 12/ 2008, domaine général.
  - 4.655.433 requêtes classées par fréquence

Besoins mis en évidence
la reconnaissance de "multi-mots"

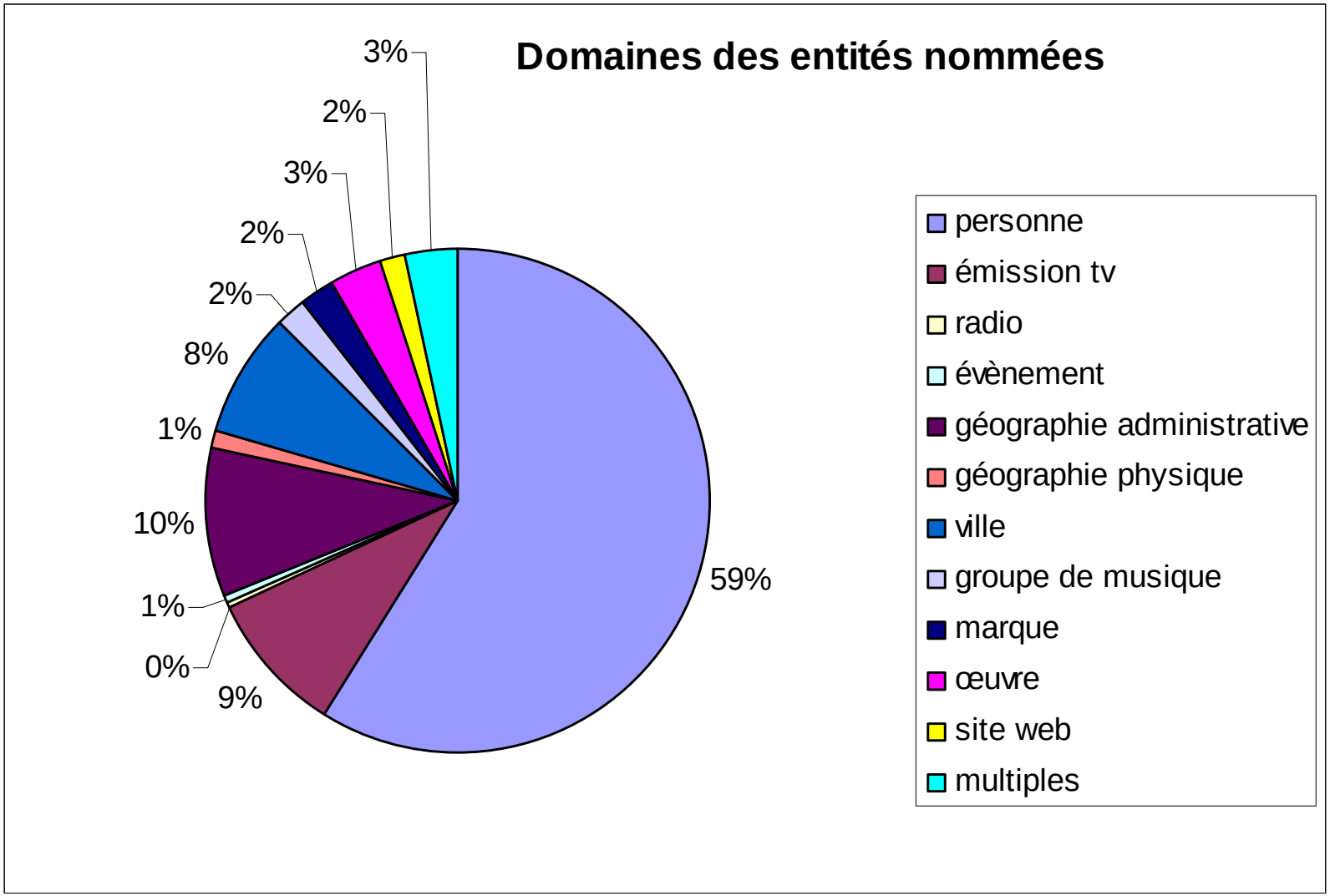
- Multi-mots
  - mots composés : *pomme de terre*, *rendez-vous, casque bleu*
  - entités nommées : noms de personnes, noms de lieux
  - termes construits : *développement durable, crack boursier, grippe aviaire*
- Usage
  - indexation : meilleure pertinence sémantique
  - requête : amélioration de la pertinence, diminution du bruit, correction contextuelle
- Multilinguisme
  - traduction spécifique
  - pas de traduction

Nombre de mots d'une requête (corpus VSE)

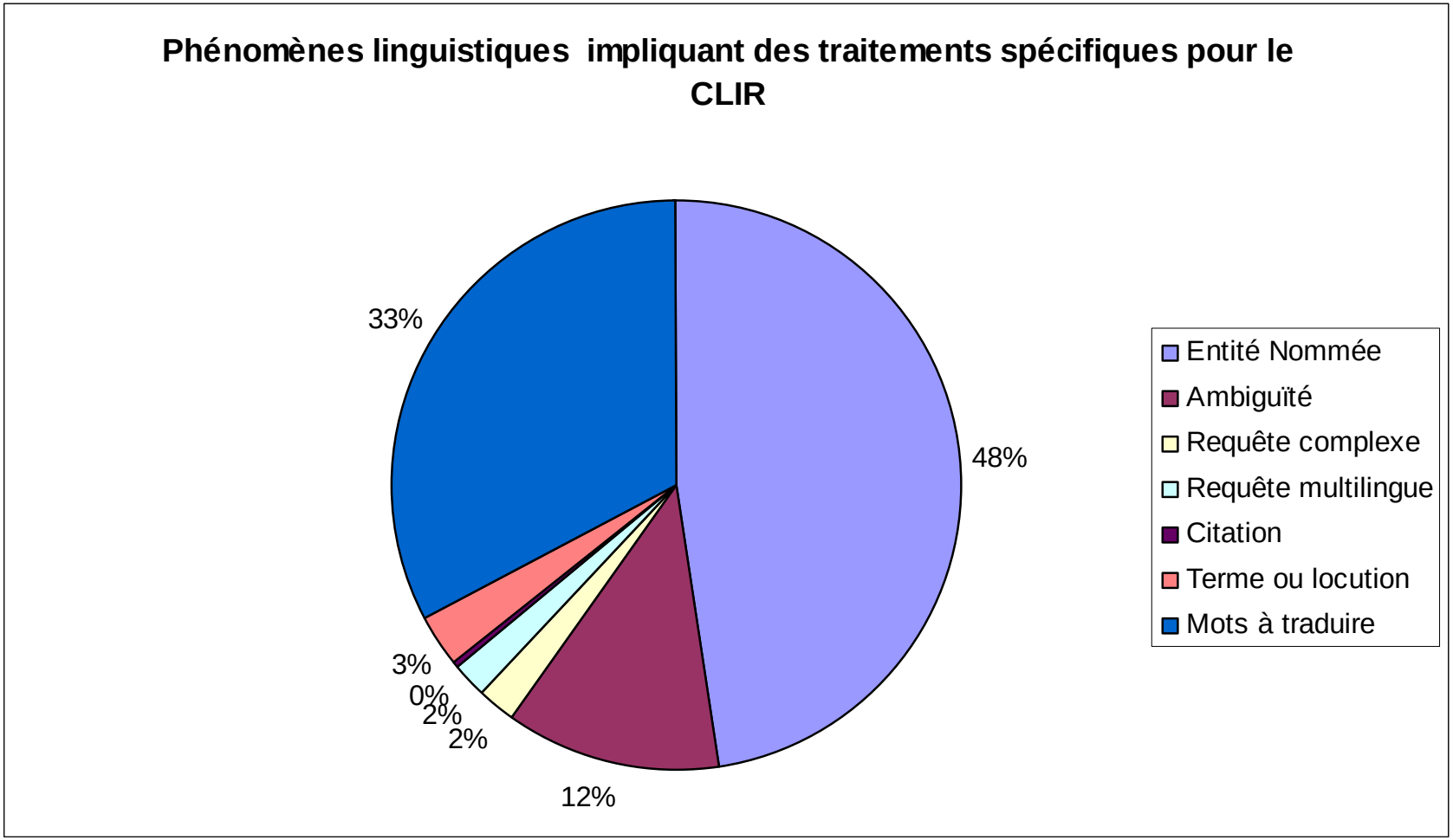


Moyenne: 1.69 mots/requête

Typologie des entités nommées contenues dans les requêtes (corpus VSE)



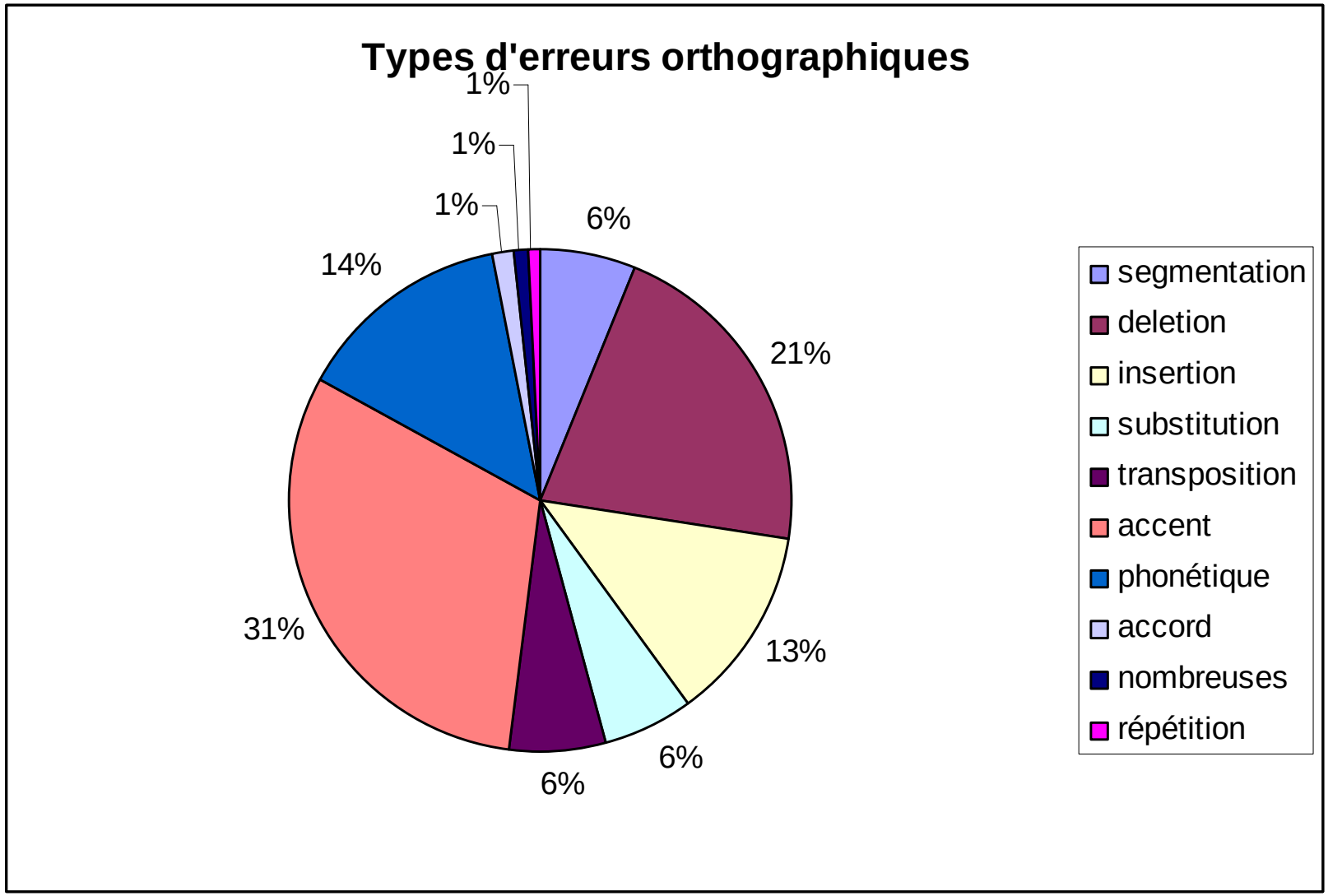
Phénomènes impliquant des traitements spécifiques pour le CLIR (corpus VSE)



Besoins mis en évidence
la correction

- Correction
  - différents types : accents, phonétique, typographique, segmentation
  - stratégie de choix : fréquence, contexte, top requêtes
- Usage
  - améliorer la pertinence : correction ou suggestion
  - complémentaire par rapport à la complétion
- Multilinguisme
  - impact sur la traduction :
    - entités nommées non reconnues
    - mots inconnus intraduisibles

Typologie des erreurs (corpus VSE)



Conclusion

- Intérêt des traitements sur le texte
  - texte omni-présent : contenus, dérivés (transcription/OCR), métadonnées
- Double usage :
  - enrichissement des contenus
  - accès aux contenus
- Besoin de traitements statistiques et de traitements linguistiques
- Besoin de ressources (lexiques, entités nommées et termes, etc.)

Sommaire

2. Travaux en Open Innovation sur le CLIR dans QUAERO-MSSE (9h45-10h30) :
   - 09h45-09h55 : Démonstration du prototype multilingue MSSE (M.Boualem & al.).

Sommaire

2. Travaux en Open Innovation sur le CLIR dans QUAERO-MSSE (9h45-10h30) :

   – 10h00-10h10 : Sources d'information et connecteurs multilingues à VSE (P.Diverres).

Sources d'information et connecteurs multilingues

- Sites de Vidéos anglais :
  - Times, Reuters, Skynews : news
  - Topgear, Skysports : sports
  - Itv : site généraliste
  - Channel4, Five : news + clips tv
  - Appletrailers : bandes annonces de films
  - Ign : bandes annonces de jeux vidéos
  - Videojug : ugc (tutoriaux en vidéos)

- Métadonnées extraites par les connecteurs : Titre, descriptif/article, date, image, Mots clefs, durée.

Sommaire

2. Travaux en Open Innovation sur le CLIR dans QUAERO-MSSE (9h45-10h30) :

- 10h10-10h20 : Architecture du prototype multilingue MSSE (Y.Almeras).

Sommaire

2. Travaux en Open Innovation sur le CLIR dans QUAERO-MSSE (9h45-10h30) :

- 10h20-10h30 : Roadmap de mise en œuvre du CLIR dans MSSE (B.Gaillard, M.Boualem).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion,
30 juin 2009

Roadmap de la mise en œuvre du multilinguisme dans le Media Search

Prototype MSSE

Projet QUAERO-MSSE et Projet VSE

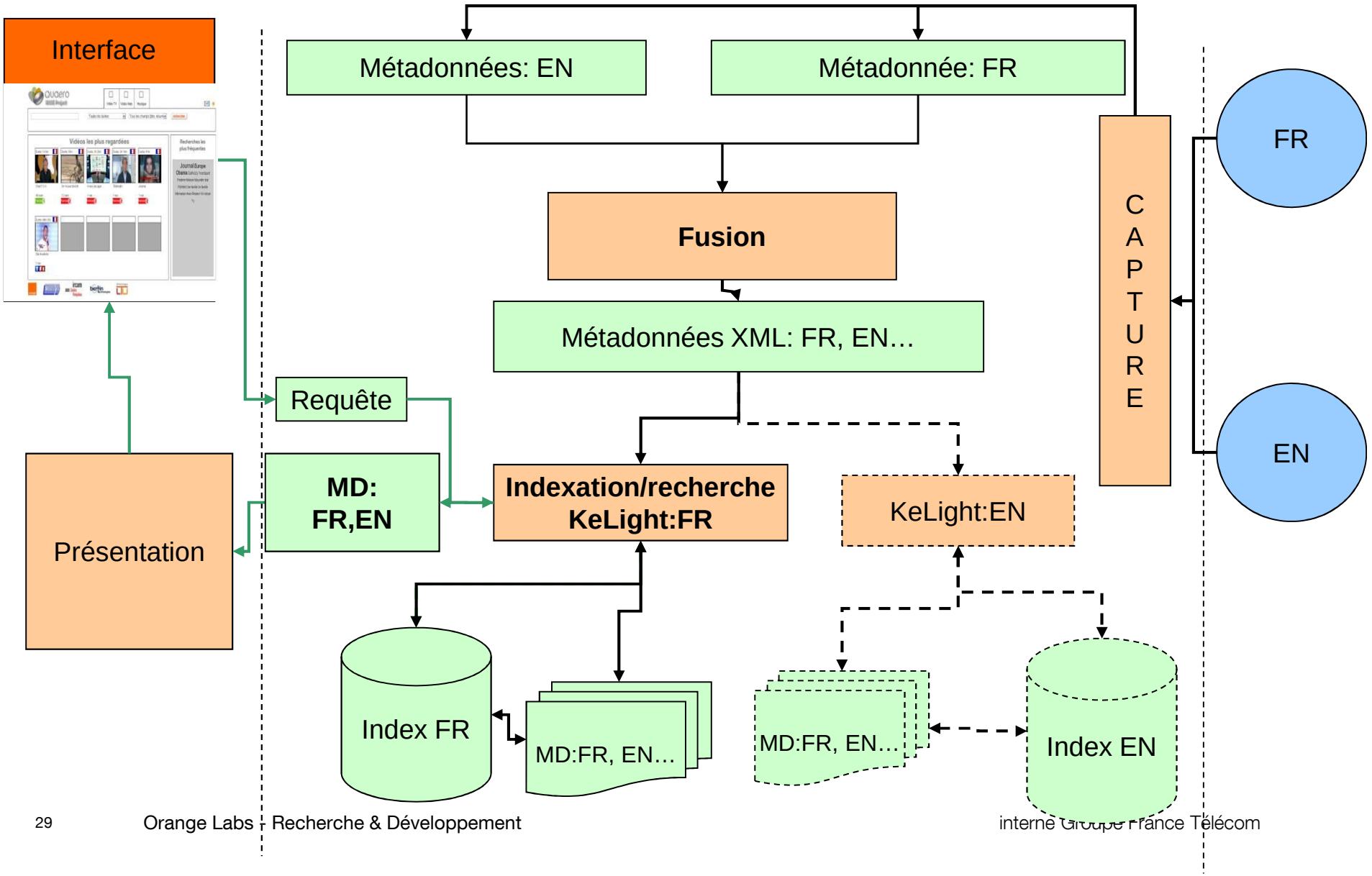
Benoît Gaillard, TECH/ACTS/FAST
Malek Boualem , TECH/ACTS/FAST

orange™

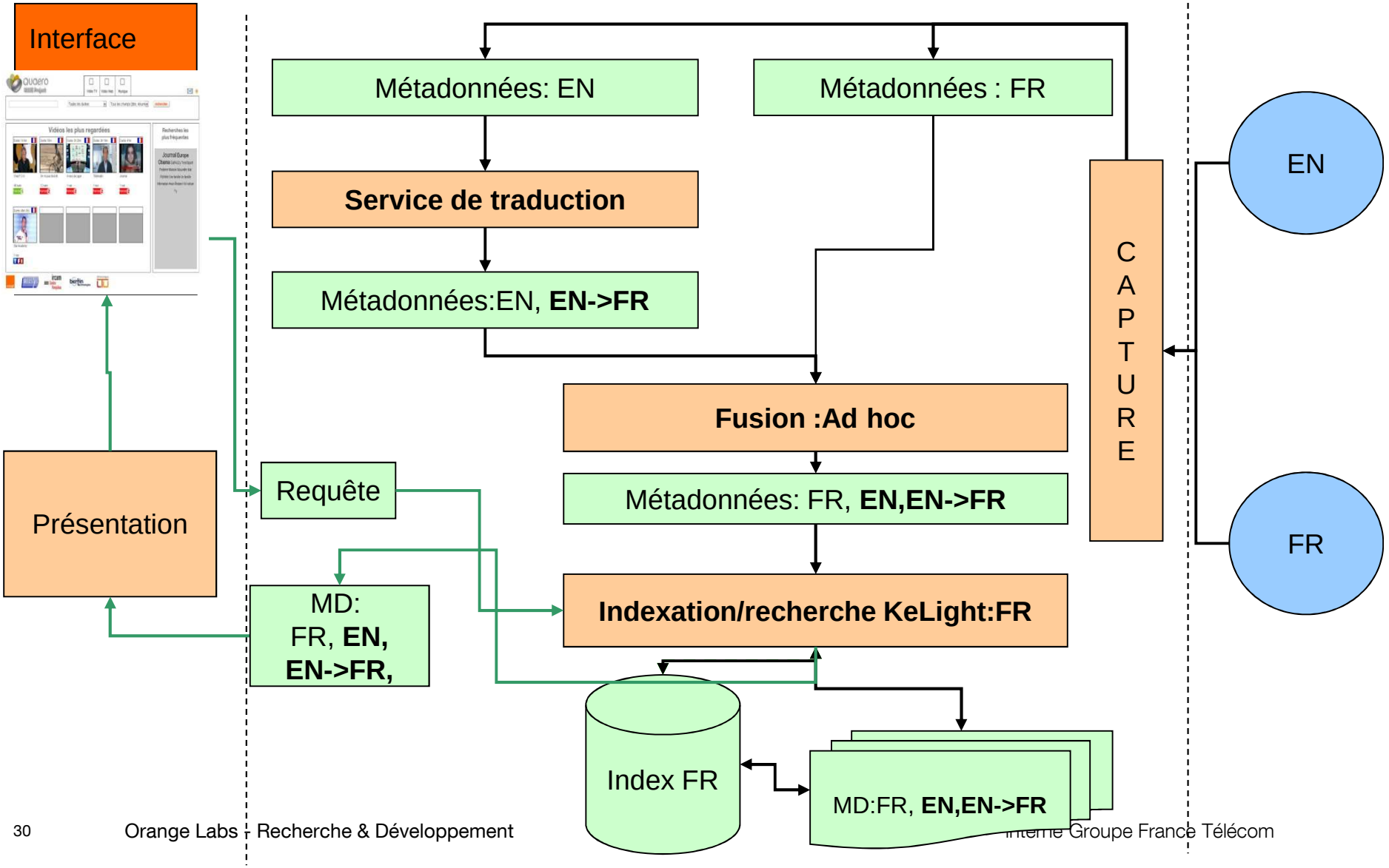
Mise en oeuvre du CLIR: Roadmap

- Plusieurs phases
  - Phase 0.1: **Recherche monolingue** dans différentes langues
  - Phase 1.0: Initialisation/**traduction des métadonnées** existantes
  - Phase 1.1: Intégration de la traduction statistique en **back-office** des métadonnées
  - Phase 2.0: Traduction statistique en **front office des requêtes** et des métadonnées
  - Phase 2.1: **Prétraitement** et traduction statistique des requêtes
  - Phase 3.0: Prétraitement et **traduction lexicale** des requêtes
  - Phase 3.1: Prétraitement et **traduction hybride optimisée** des requêtes

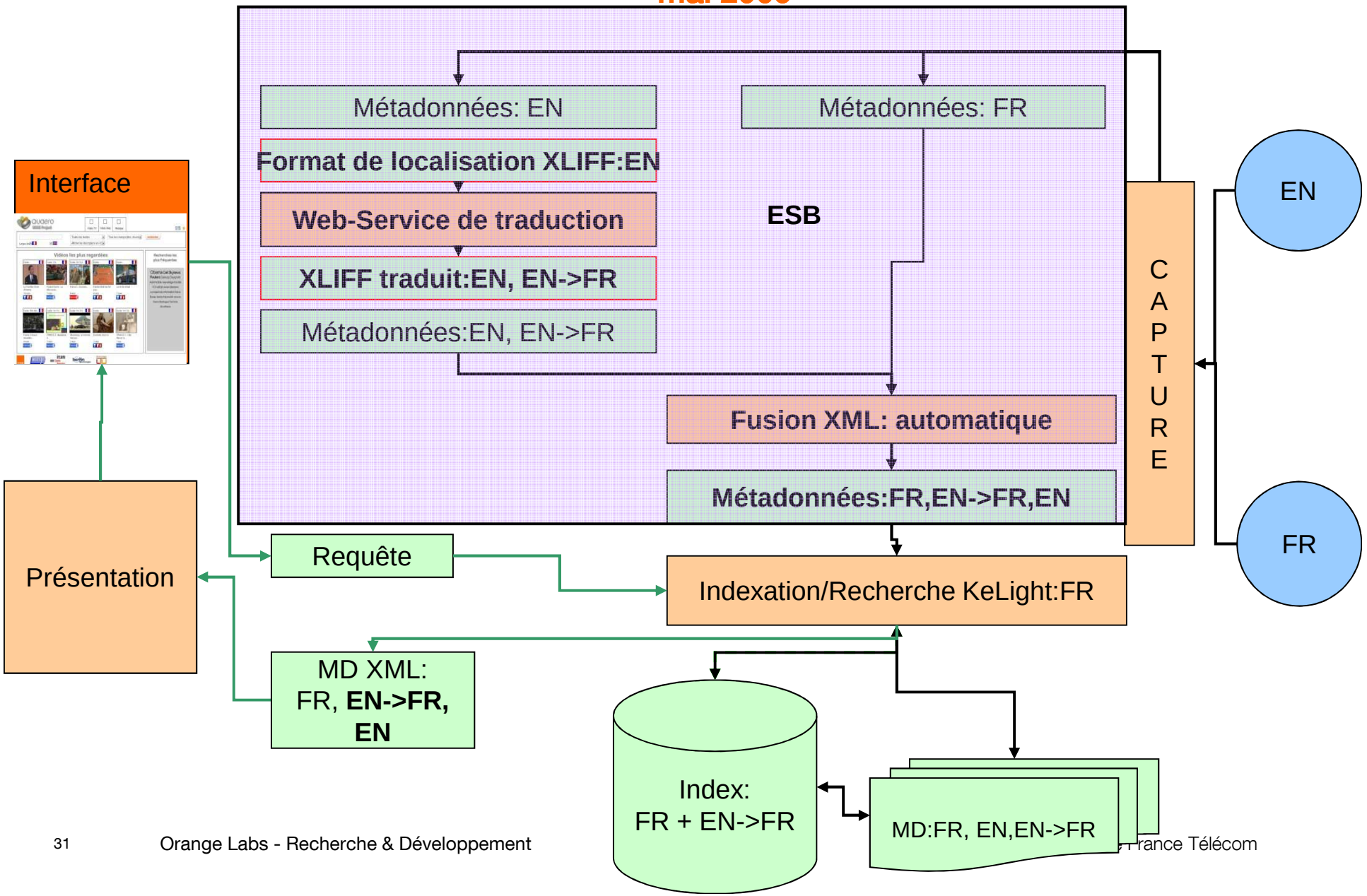
Recherche monolingue- Indexation et Recherche
- mars 2009 -



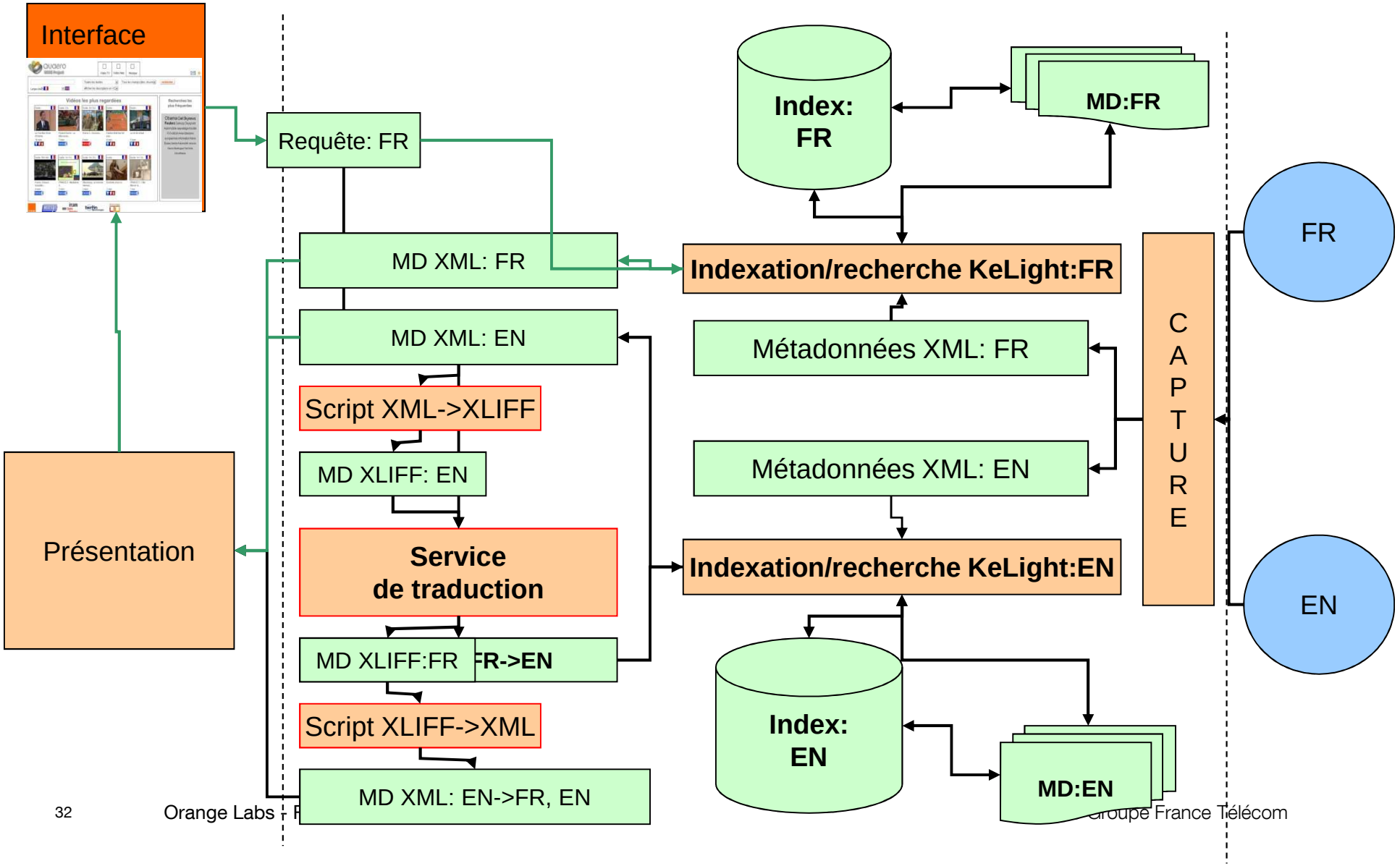
Traduction statistique des métadonnées: phase d'initialisation - avril 2009 -



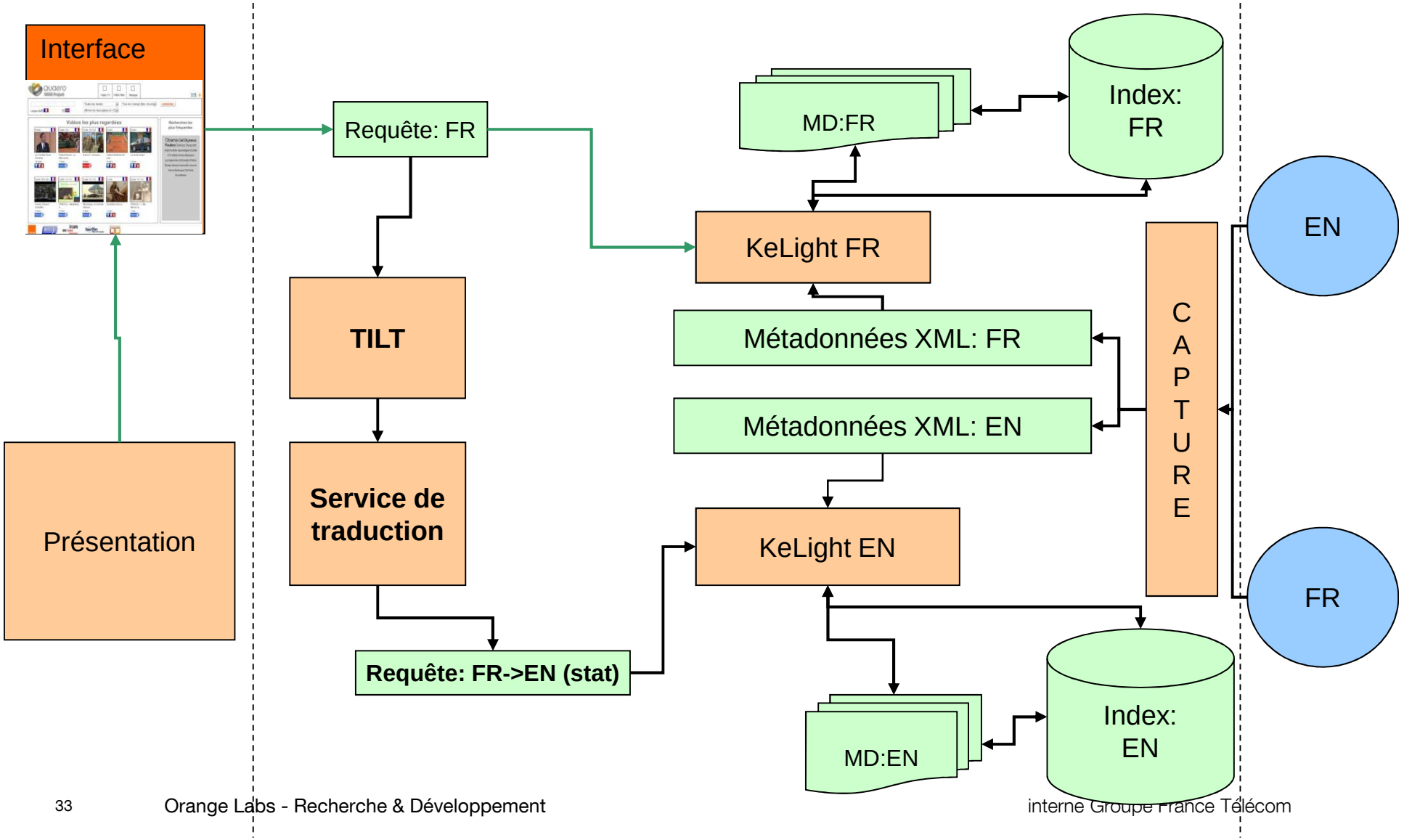
Intégration du service de traduction statistique en back-office des métadonnées
- mai 2009 -



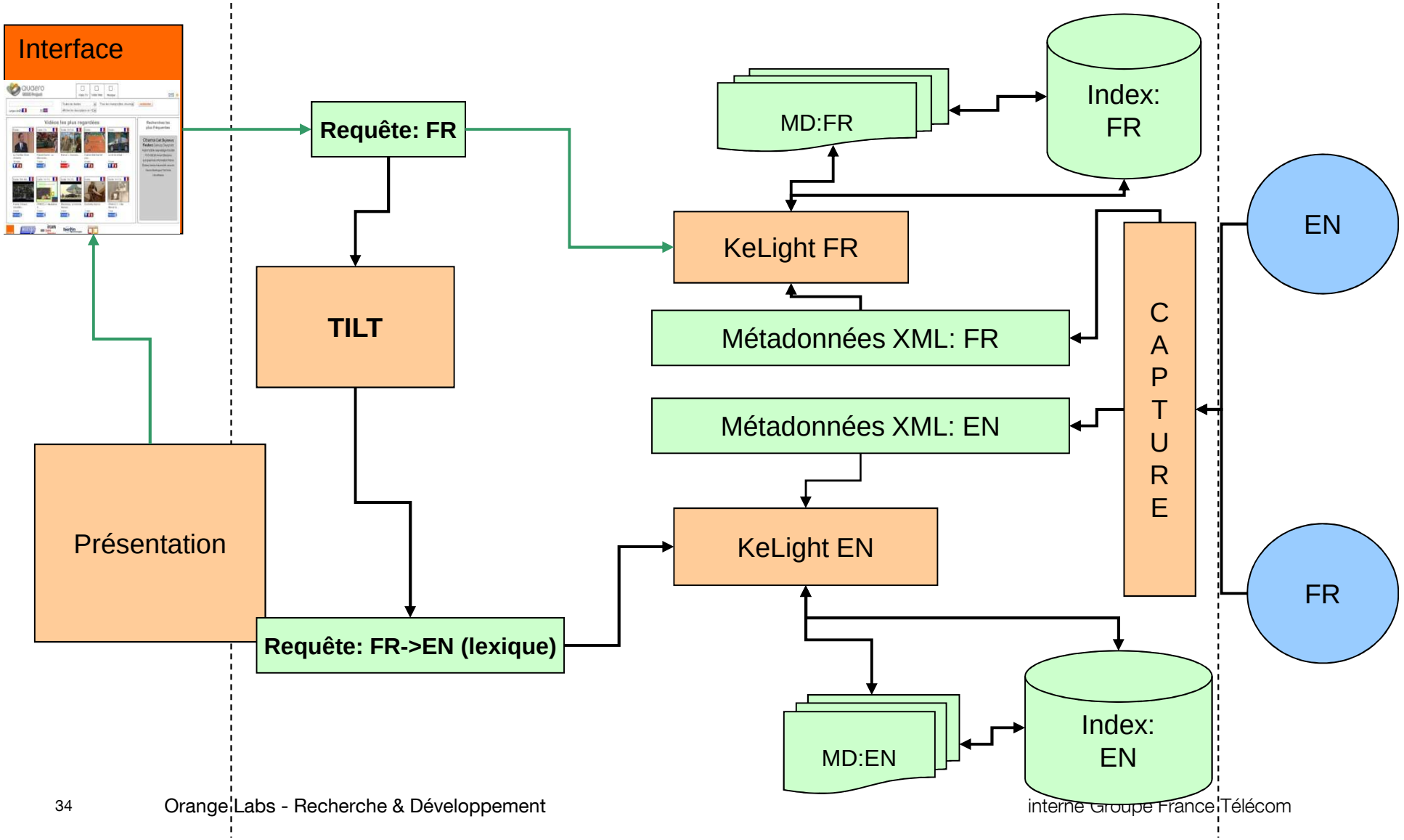
Traduction statistique en front-office des requêtes et métadonnées
- septembre 2009 -



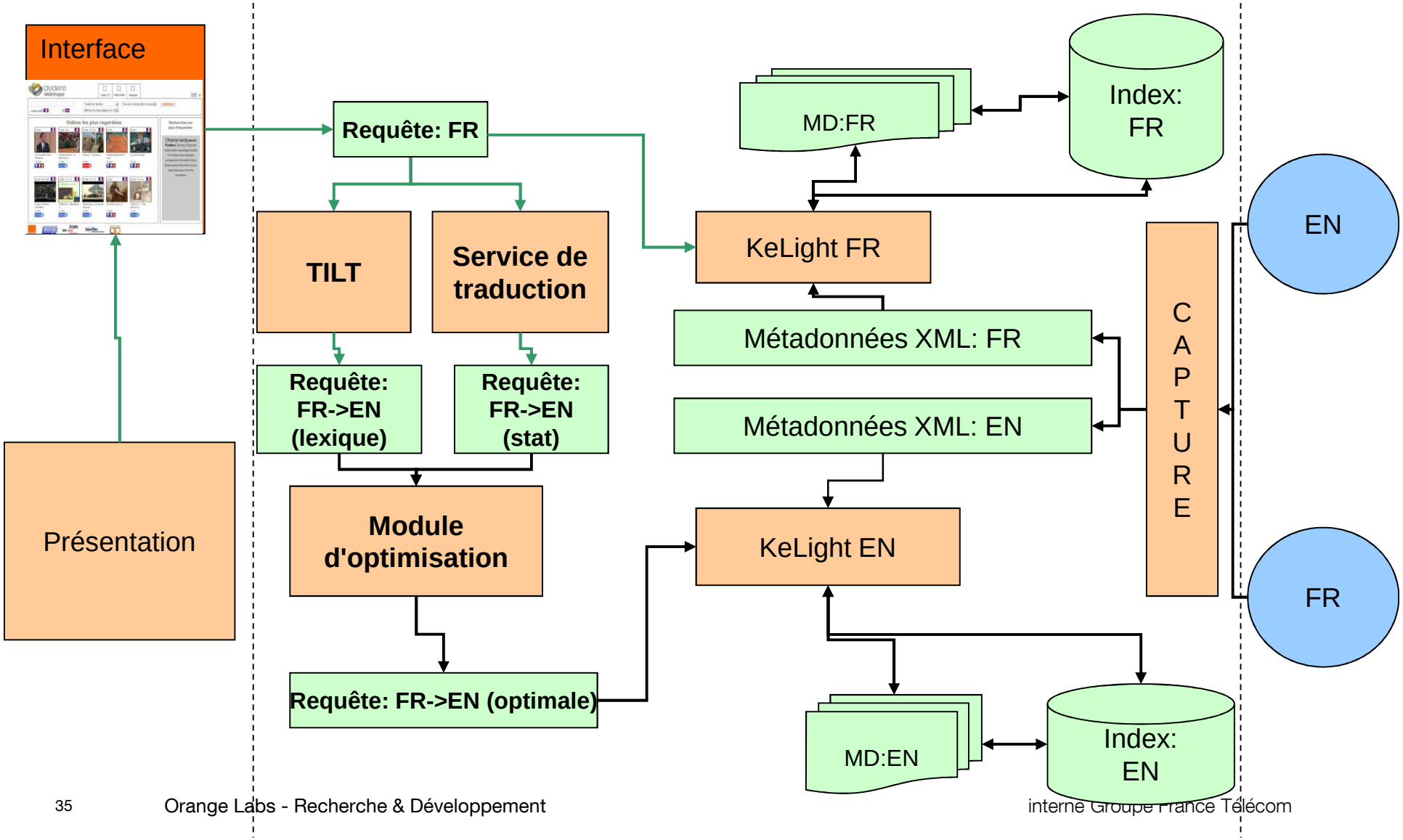
Prétraitement et traduction statistique des requêtes
- octobre 2009 -



Prétraitement et traduction lexicale des requêtes
- décembre 2009 -



Prétraitement et Traduction hybride optimisée des requêtes
- S1 2010 -



Sommaire

3. Evaluation, besoins et usages du multilinguisme (10h30-11h00) :
   - 10h30-10h40 : Evaluation de moteurs intégrant une dimension multilingue (B.Gaillard).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion,
30 juin 2009

Evaluation de moteurs intégrant une dimension multilingue

Projet VSE

Benoît Gaillard, TECH/ACTS/FAST
Malek Boualem , TECH/ACTS/FAST

Multilinguisme en Recherche d'Information: Acteurs

- Plusieurs moteurs de recherche proposent du multilinguisme
  - Google: http://www.google.fr/language_tools?hl=fr
  - Yahoo!: http://fr.docs.yahoo.com/translator/
  - Microsoft: http://www.bing.com/ http://www.fastsearch.com/
  - Youtube: http://www.youtube.com/
  - Lingway: http://www.lingway.com/ (en partenariat)
  - Xerox/Lirix: www.xrce.xerox.com/programs/lirix/index.html
  - Multimatch: www.multimatch.eu
  - PanImages: http://www.panimages.org/
  - …

Cahier de tests

- 43 tests fonctionnels issus de:
  - Etude des besoins
  - Etat de l'Art Théorique
  - Spécifications MSSE
  - Analyse de corpus de requêtes
- Catégories des tests:
  - Dimension multilingue
  - Fonctionnalités de CLIR
  - Aspects ergonomiques
  - Usage de la traduction

Synthèse & Grandes lignes de l'Etat de l'Art

- Typologie des moteurs:
  - Commercial ou prototype de recherche
- Simplicité vs qualité:
  - Transparence vs précision
  - Traduction Automatique vs précision et rappel
- Baseline et Best practice
  - Localisation. Langues > 3
  - Métadonnées associées aux images et aux vidéos
- Innovations et technologies spécifiques
  - Utilisation et constitution de lexiques multilingues
  - CLIR en mode intégré par représentation et indexations sémantique
- Insuffisances
  - Traitements linguistiques des requêtes
  - Visibilité du CLIR pour les utilisateurs

Conclusion: Recommandations et Perspectives:

- Baseline
  - Intégration de technologies de traduction automatique
  - Localisation, plusieurs langues
- Apport des traitements linguistiques
  - Traitements linguistiques pour la précision, le rappel
  - Exploitation de lexiques, thesaurus, …
- Nouveaux Usages
  - Expliciter l'intérêt du CLIR auprès des utilisateurs
  - Initier progressivement les utilisateurs à la pratique de la recherche multilingue

Livrables en lien avec cette présentation

- *Etat de l'Art et Evaluation de la Recherche d'Information Crosslingue et Multilingue (CLIR)*
- *Cahier de test pour l'évaluation de moteurs de recherche multilingues*
- *Evaluation fonctionnelle de la recherche multilingue du prototype MSSE (Cycle 1)*
- *Analyse de plusieurs corpus de requêtes des moteurs d'Orange*

Sommaire

3. Evaluation, besoins et usages du multilinguisme (10h30-11h00) :

- 10h45-10h55 : Besoins et usages du multilinguisme dans les projets d'Orange (M.Boualem).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion,
30 juin 2009

Besoins et usages du multilinguisme dans les projets d'Orange Labs

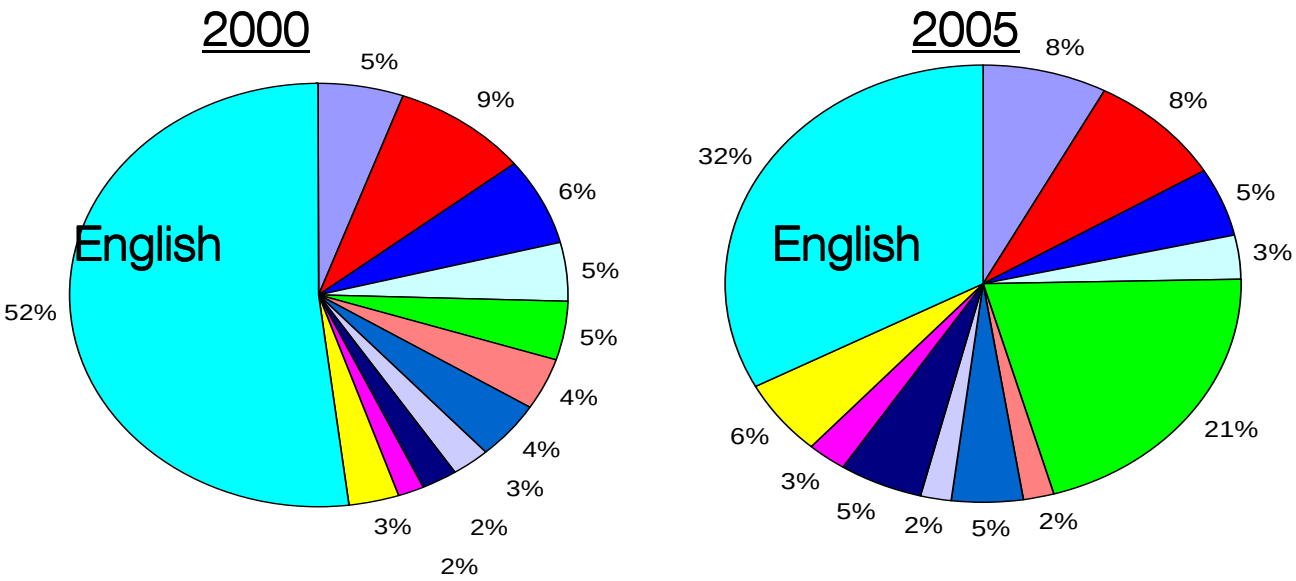
Projet VSE

Benoît Gaillard, TECH/ACTS/FAST
Malek Boualem , TECH/ACTS/FAST

orange™

Analyse générale des besoins et usages en multilinguisme pour la recherche d'information

- Diminution de l'influence de l'anglais



Analyse générale des besoins et usages en multilinguisme pour la recherche d'information

Langue	Pourcentage d'utilisateurs	Croissance relative (2000-2008)	Taux de pénétration
Anglais	29,4 %	203,5 %	21,0 %
Chinois	18,9 %	755,1 %	20,2 %
Espagnol	8,5 %	405,3 %	27,6 %
Japonais	6,4 %	99,7 %	73,8 %
Français	4,7 %	458,7 %	16,6 %
Allemand	4,2 %	121,0 %	63,5 %
Arabe	4,1 %	2063 %	16,8 %

- *Taux de pénétration* : nombre d'internautes d'une langue donnée par rapport au nombre total de locuteurs de cette langue.
- Référence : [Internetworldstats 2009].

Analyse générale des besoins et usages en multilinguisme pour la recherche d'information

- Intérêt du CLIR
  - Réduction du silence en RI
  - Intérêt pour les domaine de la veille stratégique et du renseignement
  - Simplicité de l'accès aux contenus dans la langue de l'utilisateur
  - …
- Adaptation aux utilisateurs
  - Typologie des usages et utilisateurs
    - dimension linguistique, technologique, professionnelle
  - Evolution rapide des usages
- Communautés d'intérêts:
  - SIGIR (Special Interest Group in Information Retrieval)
  - Workshop SIGIR dédié au CLIR
  - CLEF (Cross Language Evaluation Forum, http://www.clef-campaign.org/)

Multilinguisme pour les projets d'Orange Labs
Interviews

- Objet de Recherche "Media Search":
  - Henri Sanson
- VSE (INVENIO)
  - Jean Philippe Cabanal
- News Cloud & 24/24 Actu
  - Laurent Frisch
- QUAERO, MSSE
  - Malek Boualem
- PHAROS
  - Michel Plu
- Mobile Search
  - Frederic Gavignet
  - Gilles Le Calvez
- Local Search
  - Gilles Prigent
- Content Lab et Media Search on TV and Convergence
  - Kristine Kuster

Multilinguisme pour les projets d'Orange Labs

- Media Search :
  - Forte croissance de la consommation vidéo en ligne
  - Le moteur de recherche devient un mode d'accès majeur aux vidéos
  - Langues : FR, EN, ES, PL
- Content Lab et Media Search on TV and Convergence
  - 2009 : étude en cours pour affiner les besoins
  - Première phase bilingue : français/anglais.
- News Cloud & 24/24 Actu
  - Clustering : rapprochement de contenus dans d'autres langues
  - Accès à des contenus multilingues par des requêtes en français
  - Traduction des métadonnées et/ou des transcriptions

Multilinguisme pour les projets d'Orange Labs

- QUAERO, MSSE
  - Prototypage du CLIR en français/anglais
  - Extension à d'autres langues : allemand et arabe
- PHAROS
  - Fonctionnalité "more like this" multilingue
  - Traduction de la parole transcrite
  - Langues: français, anglais, allemand
- Mobile Search
  - Géolocalisation.
  - Besoins spécifiques en CLIR pour les petits pays (peu de contenu)
  - Enablers : anglais, suisse, roumain, belge.
- Local Search
  - Requêtes en langue étrangère pour des contenus locaux français (OneBox, Live Translation)
  - Géolocalisation

Synthèse

- Général:
  - Forte croissance de langues "exotiques" : Arabe, Chinois…
  - Nécessité d'adapter les sytèmes de CLIR aux cas d'usage
  - Nécessité de susciter l'attrait des utilisateurs:
    - Simplicité
    - Réduction du silence

- Orange:
  - Dimension internationale: nombreuses langues
  - Géolocalisation services pratiques (annuaires, tourisme, …)
    - À l'étranger pour des français
    - En France pour des étrangers
  - Multimedia

Livrables en lien avec cette présentation

- *Besoins et Usages de la Recherche d'Information Multilingue*

Références

- [Peters 2005] Carol Peters, "*Multilingual Information Access for digital Libraries*", www.theeuropeanlibrary.org/portal/organisation/cooperation/delos/20050915/carol_peters.ppt
- [Internetworldstats 2009]. *www.internetworldstats.com/stats7.htm*.
- [Gey et al. 2006] Fredric C. Gey, Noriko Kando, Chin-Yew Lin and Carol Peters, "*New Direction in Multilingual Information Access*", in "*ACM SIGIR Forum*", Volume 40, Issue 2, Publisher: ACM New York, NY, USA, 2006.

Sommaire

4. Traitement des requêtes pour l'accès aux contenus multilingues (11h00-11h40) :
   - 11h00-11h10 : Pré-traitement et traduction des requêtes (B.Gaillard, M.Boualem).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion
30 juin 2009

Traitement des requêtes pour l'accès aux contenus multilingues

VSE, QUAERO/MSSE

Benoît Gaillard, TECH/ACTS/FAST

Malek Boualem , TECH/ACTS/FAST

orange™

Intérêt du traitement des requêtes pour l'accès aux contenus multilingues

- Limites de la traduction systématique des contenus en back office
  - Augmentation du volume des contenus et du nombre de langues
    - Traductions non utilisées, paires de langues peu exploitées
    - Coût de stockage des traductions
    - Montée en charge du service de traduction
    - Coût économique de la traduction
  - Couverture lexicale réduite (métadonnées issues de la traduction)

➔Prétraitement et traduction des requêtes:

  - Cf limites ci-dessus
  - Aspect dynamique
  - Visibilité du CLIR par l'utilisateur

Exploration de méthodes collaboratives pour le traitement des requêtes en vue du CLIR

- Filtrage des éléments de la requête
  - Eléments traduisibles, translitérables, invariants
- Exploitation de ressources linguistiques
  - Lexiques bilingues, thésaurus, …
- Post-traitements sur la requête traduite
  - N-best, post-édition, …
- Exploitation de mesures statistiques
  - Web, corpus, …
- Exploitation de Wikipédia
  - Catégories et traductions
- Expansion sémantique
  - Compensation de la couverture lexicale réduite issue de la traduction

Filtrage des composants de la requête avant traduction

- Typologie des phénomènes
  - Locutions
  - Termes
  - Entités Nommées
  - Collocations
  - etc.

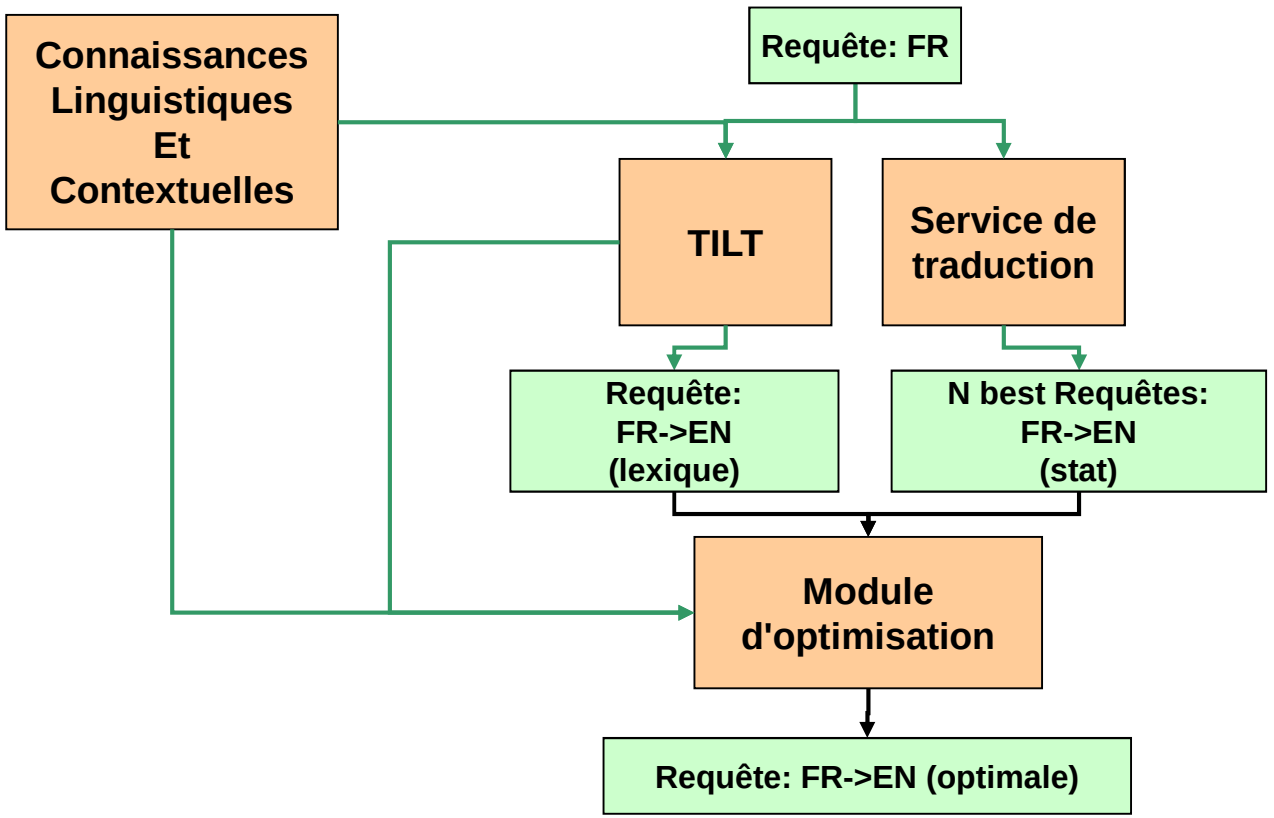
- Typologie des traitements des divers éléments de la requête
  - Eléments traduisibles
  - Eléments translitérables
  - Eléments invariants

Exploitation de ressources linguistiques

- Multi Mots dans les requêtes:
  - Corpus de requêtes OPF analysé pour la détection d'"expessions figées" (locutions, termes, mot composés, entité nommées…) .
- Identification des Entités Nommées
- Traduction d'Unités Lexicales Complexes:
  - Locutions (ou termes complexes) : appareil ménager
  - Collocations (expressions semi-figées): café noir
  - Référence : thèse (Léon-Véronis)

Post-traitements des requêtes traduites:

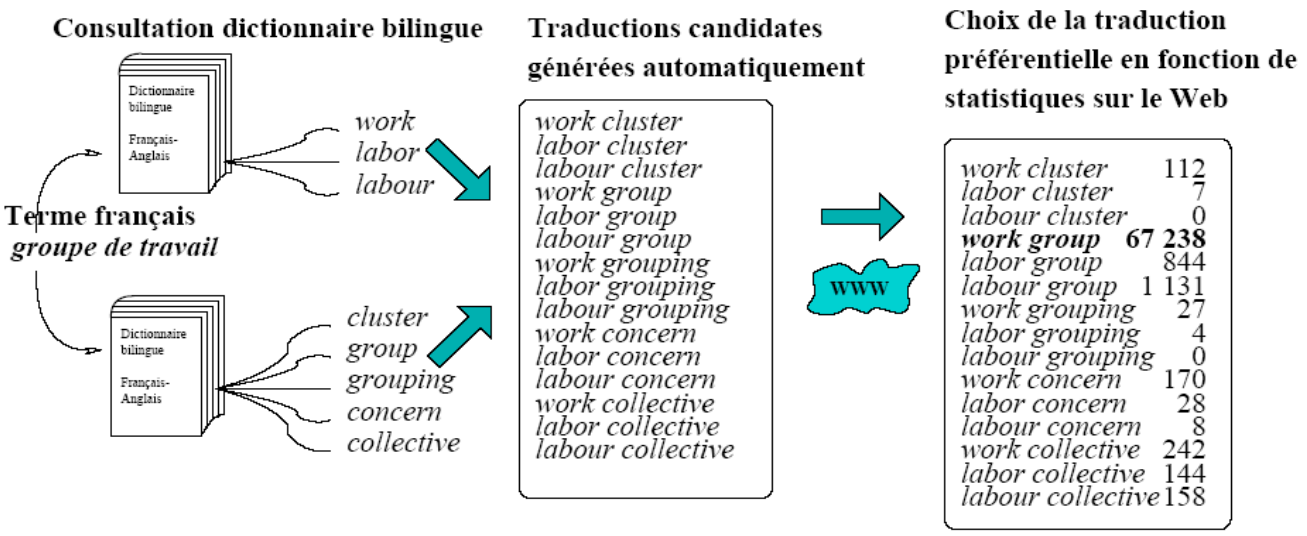
– Choix de la meilleure option/combinaison des "n-best" à l'aide de connaissances linguistiques et contextuelles:



Exploitation de mesures statistiques pour résoudre des ambiguïtés

- Peu de contexte syntaxique ou sémantique

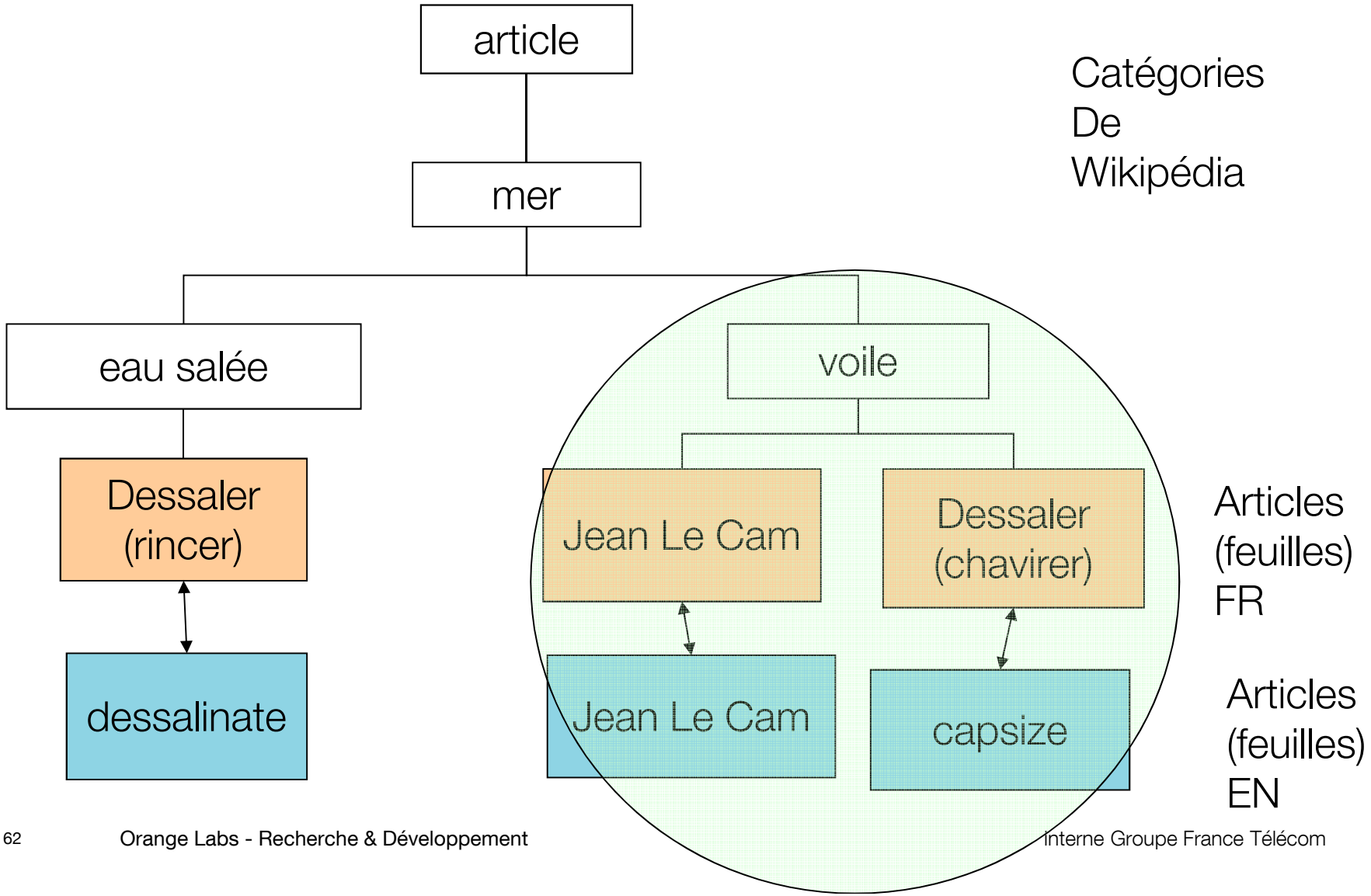
Recherche d'information interlangue



emin & Pierre Zweigenbaum, LIMSI-CNRS & SIM/DSI/AP-HP+U.P.6 École I3

Exploitation des catégories et traduction de wikipedia

– Exemple: "*dessalage Jean Le Cam*"



Exploitation de techniques d'expansion des requêtes en vue du multilinguisme:

- Compensation du caractère réduit de la couverture lexicale des métadonnées traduites automatiquement.

- Collaboration avec Jean Léon Bouraoui

Sommaire

4. Traitement des requêtes pour l'accès aux contenus multilingues (11h00-11h40) :

   – 11h15-11h25 : Expansion des requêtes pour le multilinguisme (J.L.Bouraoui, E.Guimier de Neef).

Séminaire sur le multilinguisme pour l'accès aux contenus
Orange Labs Lannion, 30 juin 2009

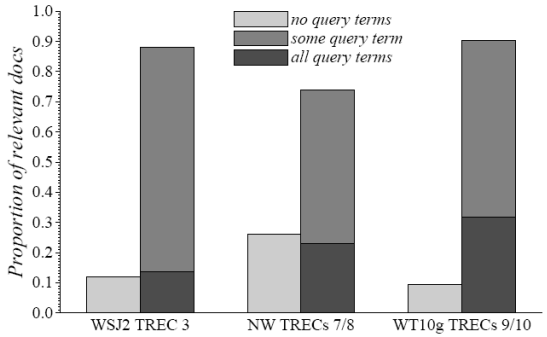
Expansion sémantique, CLIR, Media Search

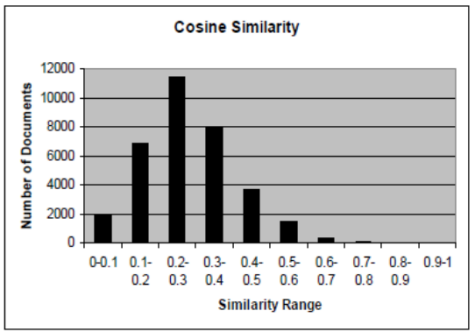
Jean-Léon Bouraoui, TECH/ACTS/FAST
Emilie Guimier De Neef, TECH/ACTS/FAST

orange™

Expansion de requête : enjeux

- En Recherche d'Information, près d'1 terme de requête sur 5 est inadéquat !
  - **20%** de mots communs à deux personnes pour désigner un même concept (études de Bates (1986) et de Furnace (1987))
  - **10 à 25%** des documents pertinents ne contiennent aucun terme d'une requête donnée (études sur *TREC*: Billerbeck (2005, p. 4))
  - **Faible similarité** entre l'espace des requêtes et l'espace des documents (Cui *et al.* (2002, p. 327))





Expansion de requête : but et méthodes

- Objectif: augmenter l'adéquation entre requêtes et documents
- Moyens: reformulation de la requête initiale et/ou ajout de nouveaux termes apparentés à celle-ci ➔ Augmentation précision/rappel
- Exemples: télévision→télé, tv, petit écran, téléviseur

  football→footballeur, entraîneur, goal

  Bretagne→Côtes-d'armor, Finistère …

Expansion de requête : valeur ajoutée au CLIR

- Liens étroits entre CLIR et expansion de requêtes
- Tous les avantages "monolingues" de l'expansion de requêtes plus:
  - Désambiguïsation (Ballesteros *et al.* (1998)). Exemple; chocolate (espagnol) → chocolate | cocoa | blood (anglais)
  - Prise en compte de nuances langagières potentiellement ignorées par la traduction (night club → discothèque, boîte de nuit, cf. Bellachia *et al.* (2008), Gaillard et Bouraoui, à paraître)

- Problématiques spécifiques:
  - Quand intervient l'expansion: avant/après la traduction?
  - Risque : Problème du "query drift" encore plus important

→ réflexion sur intégration dans l'architecture du CLIR

Quelques références…

- Ballesteros L., Croft W. B. "Resolving ambiguity for cross language retrieval". *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 6471, 1998.
- Billerbeck B., *Efficient Query Expansion*, PhD Thesis, Melbourne, Australie, September 2005.
- Bouraoui J.-L., *Etat de l'art sur l'expansion de requêtes*, Rapport interne Orange Labs, 2009
- Efthimiadis E., "Query Expansion", Williams, Martha E., ed. *Annual Review of Information Systems and Technology (ARIST)*, v31, pp 121-187, 1996.

Sommaire

4. Traitement des requêtes pour l'accès aux contenus multilingues (11h00-11h40) :

  – 11h30-11h40 : Interprétation des requêtes à base d'ontologies (J.Heinecke, E.Guimier de Neef).

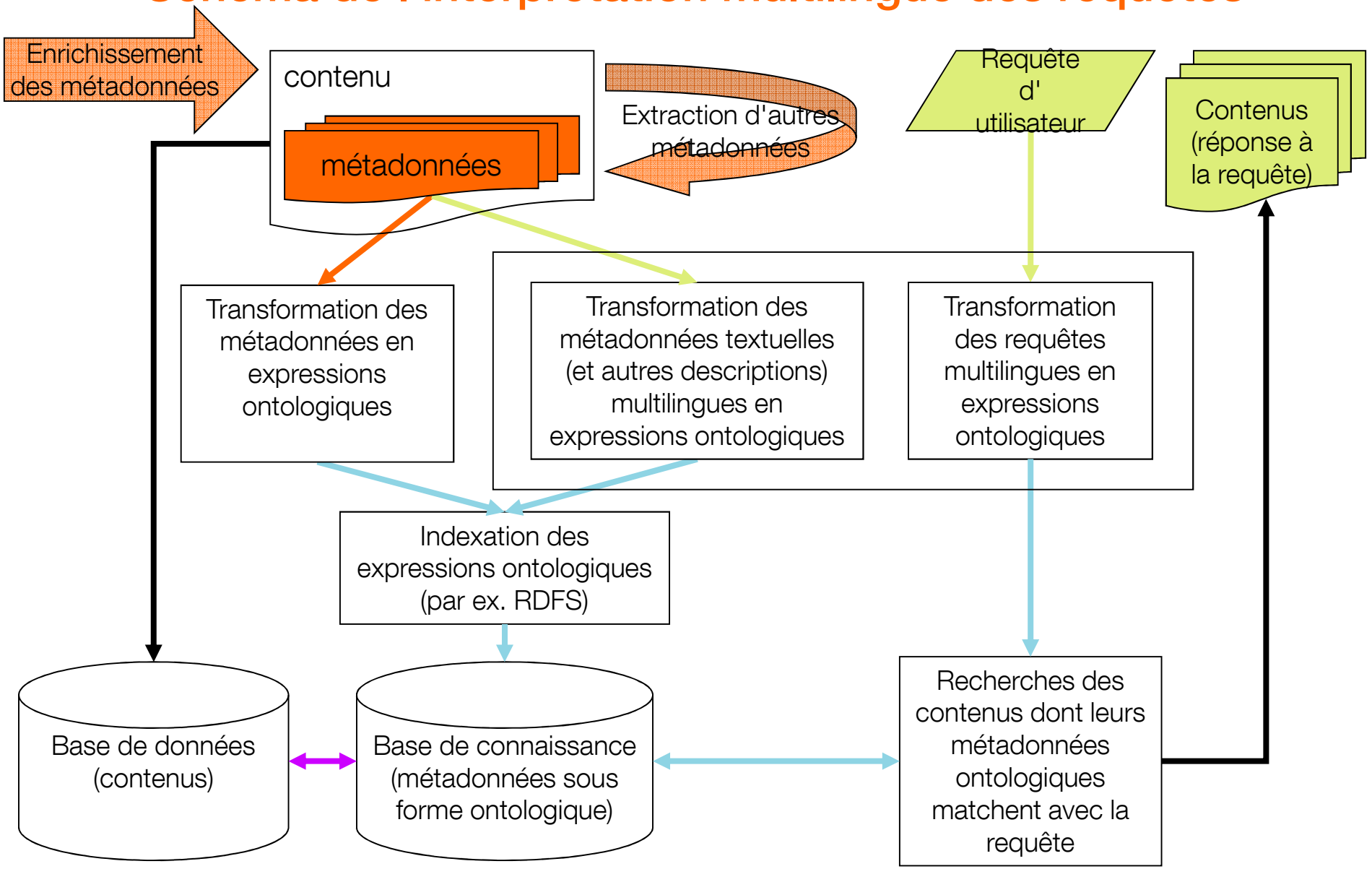
Interprétation multilingue des requêtes à base d'ontologies

Orange Labs

Johannes Heinecke, Recherche & Développement
30 juin 2009, Séminaire sur le multilinguisme pour l'accès aux contenus

orange™

Schéma de l'interprétation multilingue des requêtes



Quitter le niveau de langue pour un niveau conceptuel

- L'objet de l'indexation et de la requête ne sont pas des textes/requêtes dans une langue ou une autre, mais des **expressions ontologiques** (par ex. RDFS ou OWL) créées à partir des textes/requêtes
  - L'indexation et le recherche sont donc totalement indépendants des langues
- L'index (→utilisé par le moteur) est une **base de connaissance**
  - Cela nécessite une **ontologie** qui couvre tous les **concepts** pertinents de l'application
  - Pour l'indexation, les contenus textuels et leurs métadonnées sont transformés en **expressions ontologiques** (par ex. RDFS ou OWL).
  - Pour chaque langue des **données linguistiques liées avec l'ontologie** doivent exister

Avantages

- Indexation et recherche dans les **données extra-linguistiques** (c-a-d ontologiques)
- On garde la **signification des relations** entre les mots
- On ne garde que les **concepts pertinents** à l'application
- Plusieurs technologies existent
  - plus robuste (avec un nombre de relations entre concepts limité)
  - moins robuste (avec le risque d'échouer sur des textes/requêtes trop agrammaticaux.

- Inconvénients
  - Nécessite des ontologies qui couvrent tous les concepts qu'on veut indexer et pouvoir utiliser dans les requêtes

Sommaire

5. Multilinguisme pour 24/24 Actu (11h45-12h00) :
   - 11h45-11h55 : Expérimentation bilingue pour 24/24 Actu (P.Filoche et T.Urvoy).

do you 24/24 ?

petite expérimentation bilingue pour 24/24 Actu

Orange Labs

pascal filoche, TECH/ACTS/FAST

tanguy urvoy, TECH/ASAP/PROF

avec le concours de jonathan chevelu, TECH/ASAP/NADIA

30/06/2009

orange™

mise en œuvre

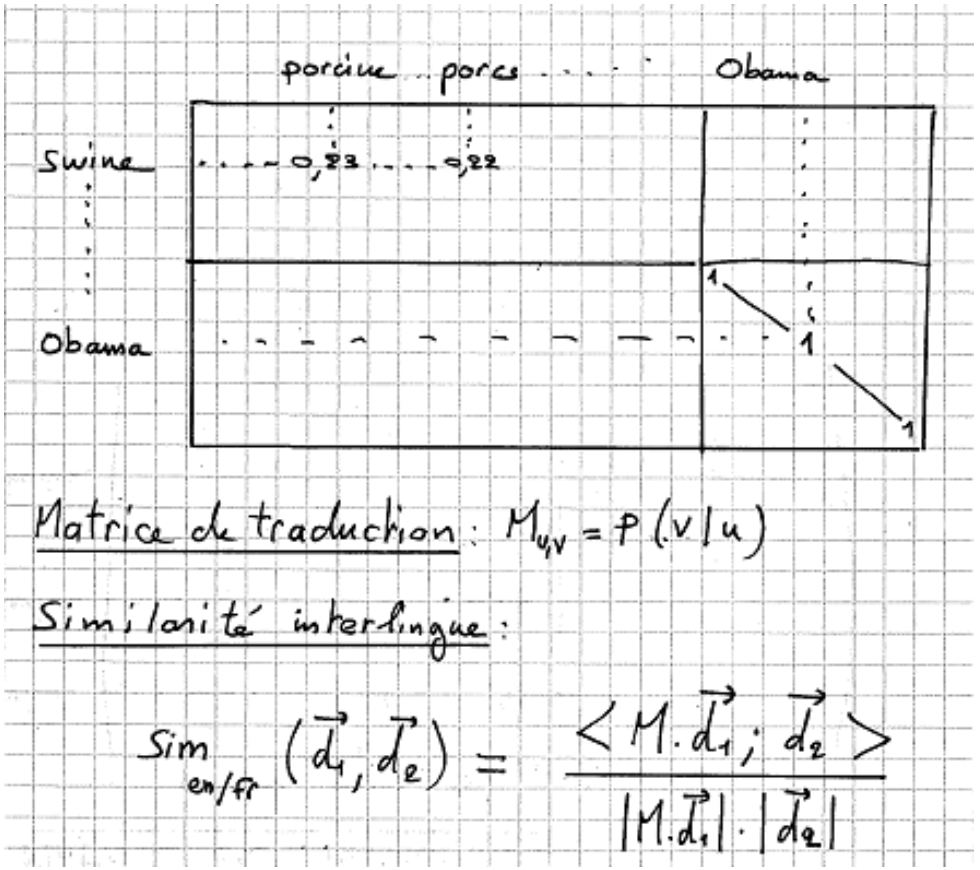
➔ comment inclure des contenus anglophones dans le service actuel ?

- réaliser une collecte de flux anglophones
- modéliser ces contenus sous une forme comparable aux données actuelles

➔ comment obtenir cette forme comparable entre français et anglais ?

proposition : mettre en œuvre une traduction statistique

english	français	probabilité
swine	porcine	0,23
swine	porcs	0,22
swine	aphteuse	0,15
swine	classique	0,13
swine	peste	0,08
swine	cuisine	0,05
swine	aviaire	0,05
swine	alimentation	0,05
swine	apparition	0,03
swine	abattage	0,03
swine	urgences	0,00
swine	transmission	0,00



➔ mais comment obtenir cette matrice de traduction ?

→ grâce au parlement européen !

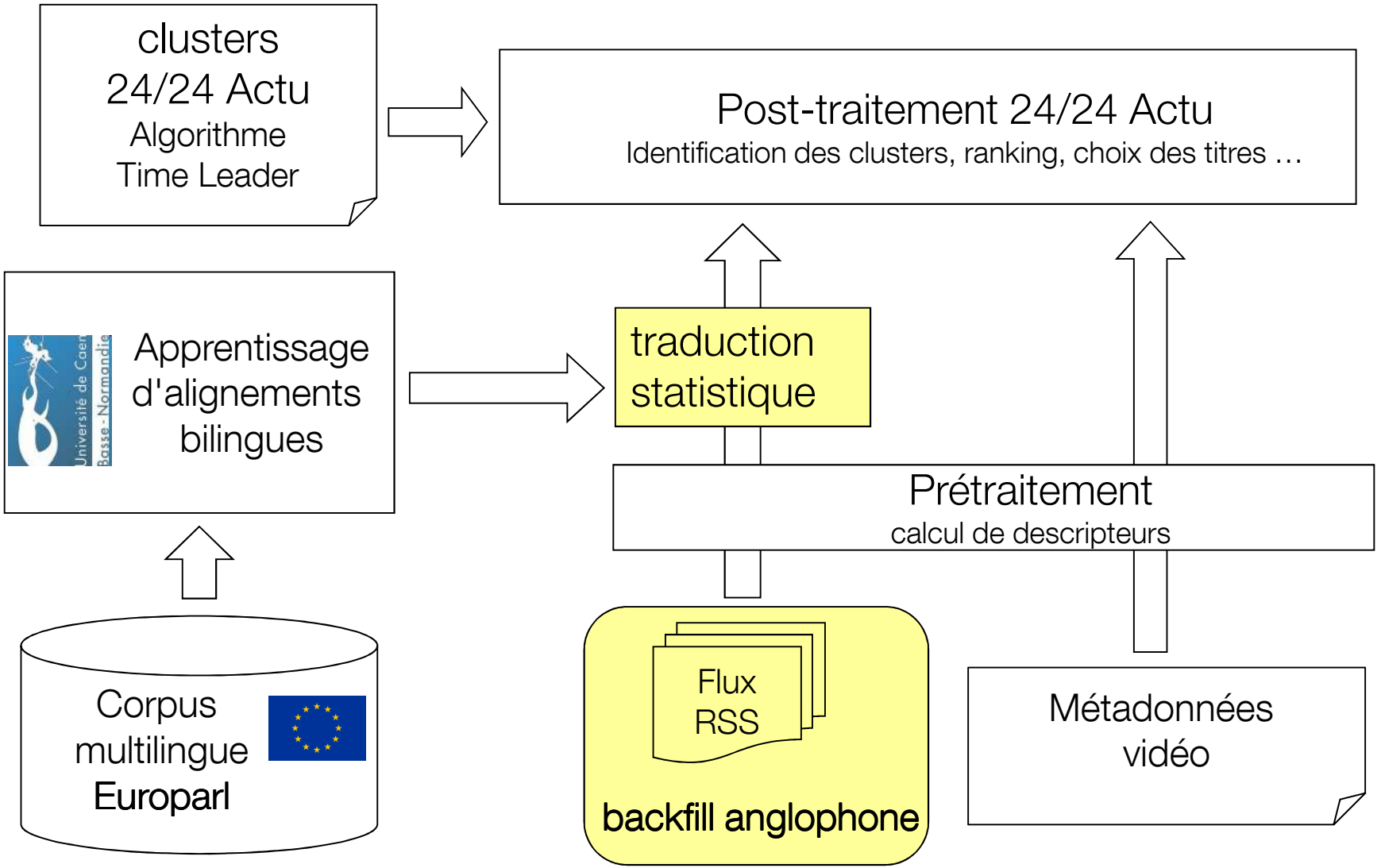
- corpus europarl : 40 millions de phrases alignées
- comptes-rendus traduits en 11 langues

(French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek, Finnish)

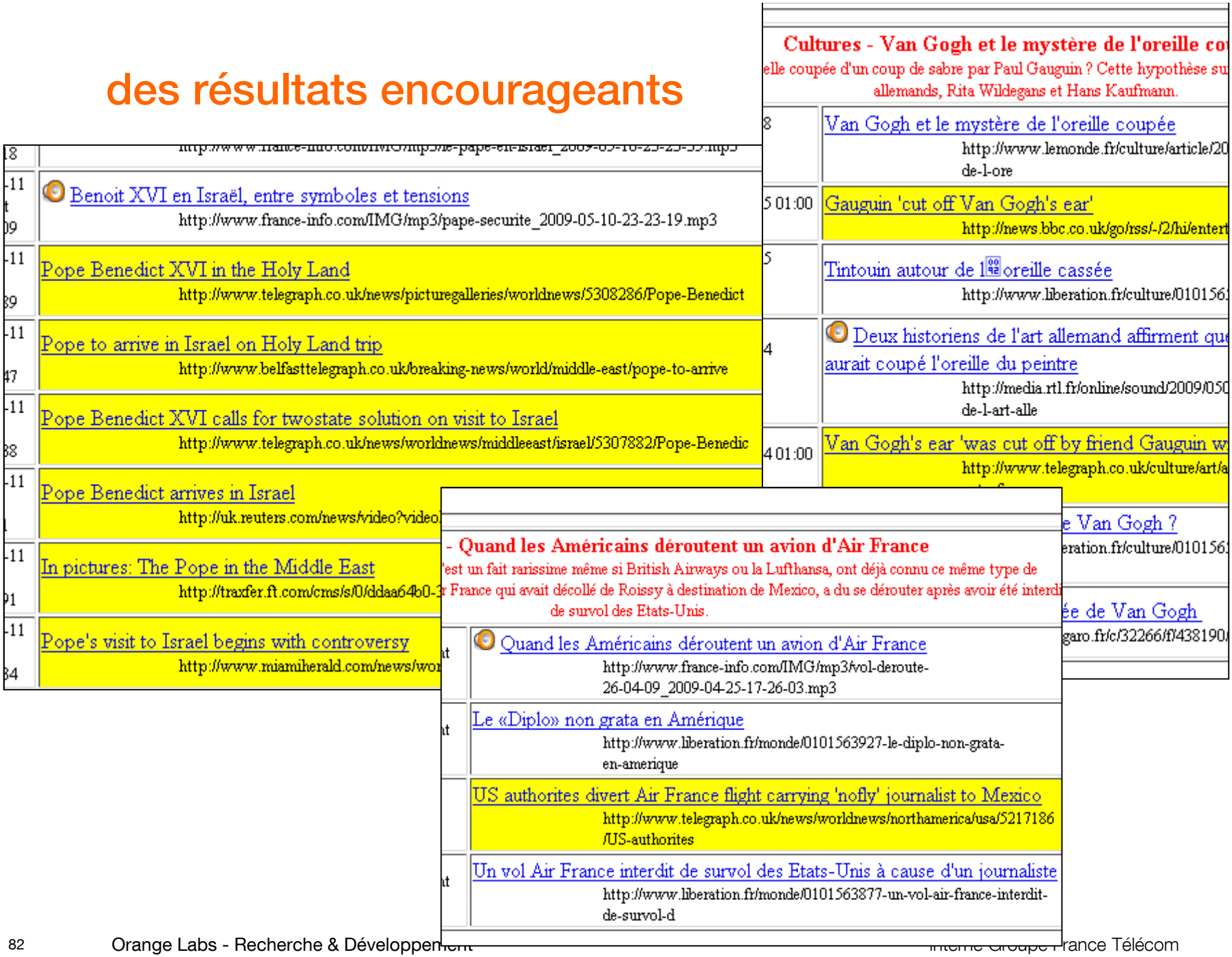
apprentissage par alignement

english	français
mr president , we are dealing here with amendments to a directive concerning the intra-community trade in bovine animals and swine .	monsieur le président , nous avons ici affaire à des amendements visant à modifier une directive relative aux problèmes de politique sanitaire en matière d' échanges intracommunautaires d' animaux des espèces bovine et porcine .
if britain is once again to reactivate a dynamic intra-community trade in bovine animals and swine , then immediate help is required to alleviate the catastrophic situation affecting rural britain and affecting our pig and beef sectors in particular .	si la grande-bretagne veut redynamiser les échanges ~~intracommunautaires des bovins et~~ porcins , elle doit immédiatement fournir une aide afin de soulager la campagne britannique de cette situation catastrophique qui touche rudement ~~les secteurs~~ porcin et bovin en particulier .
we had the bse crisis in europe and several european countries suffered from swine ~~fever outbreaks~~ .	nous avons connu en europe la crise de l' esb . nous avons ~~connu la peste~~ porcine dans différents pays européens .
i am dealing with four reports relating to the dairy sector , the cereals sector and swine ~~fever involving total~~ spending of approximately eur 16 billion , which is a considerable proportion of the budget .	je m' occupe de quatre rapports concernant les secteurs ~~laitiers , des céréales et de la peste~~ porcine dont les dépenses s' élèvent environ à 16 milliards d' euros , ce qui constitue une part considérable du budget .
they cover the following subjects : milk powder , butterfat , cereals and swine fever .	ils abordent les domaines du lait en poudre , des matières grasses butyriques , des céréales et de la peste porcine .
very briefly on the subject of swine fever : the commission is of the same mind as those who feel that after the last epidemic , there is an urgent need to revise the relevant legal provisions .	très brièvement sur la sujet de la peste porcine : la commission estime aussi qu' il y a maintenant un besoin urgent , après la dernière épidémie , de retravailler les dispositions juridiques en la matière .
when we had swine fever , we gave eur 800 million to german and dutch farmers .	à l' époque de la fièvre porcine , nous avons donné 800 millions d' euros aux agriculteurs allemands et néerlandais .

architecture logique



des résultats encourageants




Ten suspected swine flu cases investigated
http://news.scotsman.com/scotland/Ten-suspected-swine-flu-cases.5258475.jp
Liverpool News: Third Merseyside case of swine flu confirmed in
http://www.liverpooldailypost.co.uk/liverpool-news/regional-
Finland confirms 2 swine flu cases
http://www.etaiwannews.com/etn/news_content.php?id=946
WHO: 30 countries report 4,694 cases of swine flu
http://www.earthtimes.org/articles/show/268241,who-30-cou
Fresh swine flu cases in UK found but none are in Scotland
http://news.scotsman.com/health/Fresh-swine-flu-cases-in.52
Fresh swine flu cases in UK found but none are in Scotland
http://thescotsman.scotsman.com/uk/Fresh-swine-flu-cases-ir
Ten more cases of swine flu in England
http://www.belfasttelegraph.co.uk/breaking-news/world/euro
La ministre de la Santé Roselyne Bachelot sur LCI : "Nous de
http://media.rtl.fr/online/sound/2009/0506/4749266_La-minis
Number of confirmed cases of swine flu down by one
http://www.radionz.co.nz/news/stories/2009/05/06/1245af7a
Etat de l'épidémie en France et moyens de lutte : la ministre de
http://media.rtl.fr/online/sound/2009/0504/4733907_La-minis
Santé - Grippe A - Bilan des cas avérés dans le monde
BFM-TV/BFM-TV_1642_2079_7hPremiereEdition_2009056
Santé - Grippe A - Bilan des cas avérés dans le monde
SRI LANKA : Paris et Londres préoccupés par le sort des civils
http://flv.france24.com/FR NW PKG SRI LANKA UPDATE .flv
Les chefs de la diplomatie française et britannique ont déploré, lundi, les difficultés d'accès au Sri Lanka pour l'ONU, qui, au lendemain de la mort de 100 enfants dans des bombardements, a dénoncé un "bain de sang".
De nombreuses mobilisations "contre le génocide des tamouls"
http://media.rtl.fr/online/sound/2009/0512/4848518_De-nombreuses-mobilisations-c
UN anger over Sri Lanka 'bloodbath' as 1,000 Tamil civilians killed by army shelling
http://thescotsman.scotsman.com/world/UN-anger-over-Sri-Lanka.5255873.jp
UN anger over Sri Lanka 'bloodbath' as 1,000 Tamil civilians killed by army shelling
http://news.scotsman.com/world/UN-anger-over-Sri-Lanka.5255873.jp
Britain, UN appalled by 'bloodbath' in Sri Lanka
http://www.radionz.co.nz/news/stories/2009/05/12/1245afd65e4b
Nouveaux bombardements meurtriers au Sri Lanka
http://www.lemonde.fr/asie-pacifique/article/2009/05/12/nouveau-bombardement-meu
ONU et ONG s'élèvent depuis plusieurs semaines contre une catastrophe humanitaire dans le nord-est de l'île
http://media.rtl.fr/online/sound/2009/0511/4798012_ONU-et-ONG-s-elevent-depuis-p
AFP
Sri Lanka: plus de 100 enfants tués, l'ONU dénonce un "bain de sang"
urn:CNG.c000400da7b1e3d8eb3063898d848468.61

même en l'absence d'entité nommée "facile"

Société - Les Françaises se voient plus grosses qu'elles ne le sont

Dorothée Barba Les Françaises, dont la corpulence est la plus faible de toute l'Europe, se voient plus gro
ude publiée mercredi dans le bulletin d'avril de l'Institut national d'études démographiques (Ined), "Popu

2009-04-23 17:16 Soc sim: 0.376	Les Françaises se voient plus grosses qu'elles ne le sont http://www.france-info.com/IMG/mp3/les-moins-gross d-europe_2009-04-23-17-15-4
2009-04-23 14:45 Soc sim: 0.555	Cette étude de l'institut d'études démographiques révèle n classement dans l'Europe http://media.rtl.fr/online/sound/2009/0423/4579669_Ce de-l-institut-d-e
2009-04-23 01:00 sim: 0.568	French women are thinnest in Europe but think they're fat http://www.telegraph.co.uk/news/worldnews/europe/fra /French-women-are
2009-04-23 00:00 Soc sim: 1	Les femmes françaises sont les plus minces http://rss.leparisien.fr/item-1675395-999886893.html

ie - Small Elite Reaps Millions in E.U. Farm Subsidies

than 700 subsidies of at least $1.33 million in 2008, according to statistics from 26 of the 27
Union nations.

- Agriculture : la PAC dévoile ses bénéficiaires (0.291 - 2009-04-30 16:21)

Small Elite Reaps Millions in E.U. Farm Subsidies
http://feeds.nytimes.com/click.phdo?i=3dadda1792e79868056b6ed

"Le journal économique" du 1er mai 2009
http://media.rtl.fr/online/sound/2009/0501/4707601_Le-journal-eco
du-1er-

La France lève le voile sur les bénéficiaires des subventions agricoles
http://www.lemonde.fr/economie/article/2009/04/30/la-france-leve-l

bénéficiaires de subventions agricoles
rss.leparisien.fr/item-828880-999518616.html

Économie - USA: les pertes d'emplois diminuent mais le chômage con

rès nettement marqué le pas en avril aux Etats-Unis, essentiellement grâce à l'Etat, mais le chômage continue de grimper et l'écon
chiffres corrigés des variations saisonnières publiés vendredi par le département du Travail, 539.000 e

:29 Eco	AFP USA: les pertes d'emplois diminuent mais le chômage continue son ascension urn:CNG.659625bdca36768d4ca4e1443505c0cd.41
:50 Eco	ÉTATS-UNIS : Le chômage grimpe à 8,9 %, les pertes d'emplois ralentissent http://flv.france24.com/FR_NW_PKG_CHOMAGE_USA_23H.flv Malgré un taux de chômage qui culmine à 8,9 % - son plus haut niveau depuis septembre 1983 -, le rythme des des économie qui reste encore faible.
:00	Unemployment up, fewer jobs lost http://www.washingtontimes.com/news/2009/may/08/unemployment-fewer-jobs-lost/
:00 Eco	USA : le chômage au plus haut depuis 1983 http://www.lefigaro.fr/economie/2009/05/08/04001-20090508ARTFIG00351-usa-le-chom
:00 Eco	Le chômage américain au plus haut depuis 1983

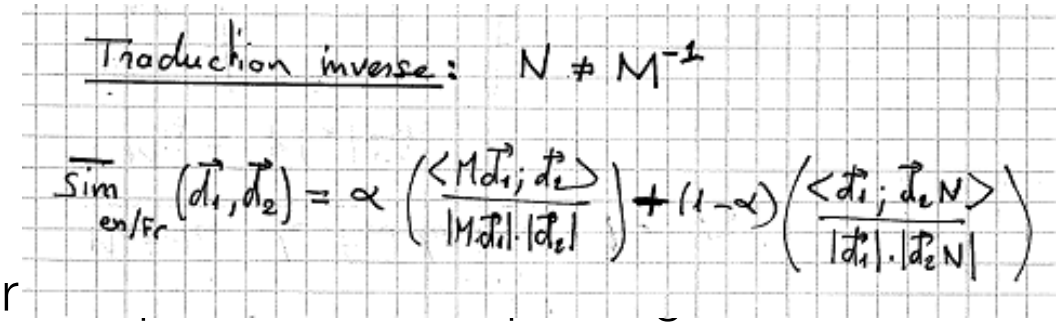
pistes de poursuite – côté service

- un onglet "dans la presse étrangère" ?
- mise en avant des pays d'origine des news

- focus sur les news relatives à la France ou aux pays concernés concerné par l'information – géolocalisation des news

pistes de poursuite – côté technique

- stabiliser la similarité interlingue
  lisser les scores à l'aide de la matrice inverse FR → EN



- Enrichir
- Réaliser un clustering anglophone complet
  et des rapprochements cluster/cluster
  (plus robuste, surtout si un 2424news.co.uk voit le jour)
- Utiliser les clusters bilingues pour améliorer les connaissances d'alignement (boucle vertueuse)

merci

orange™