

Conception et prototypage d'un outil web de médiation et d'aide au dialogue tchaté écrit en langue seconde

Thèse soutenue par:
- Achille Falaise

Sous la direction de:
- Christian Boitet
- Hervé Blanchon

Société Prosodie SA

Laboratoire LIG

Équipe GETALP

Université Joseph Fourier

Introduction

Objectif

- Aides au dialogue en langue seconde ($\neq$ langue maternelle)
 - pour des locuteurs non-natifs maîtrisant mal la langue...
 - ... et pour leurs interlocuteurs natifs

Problématique

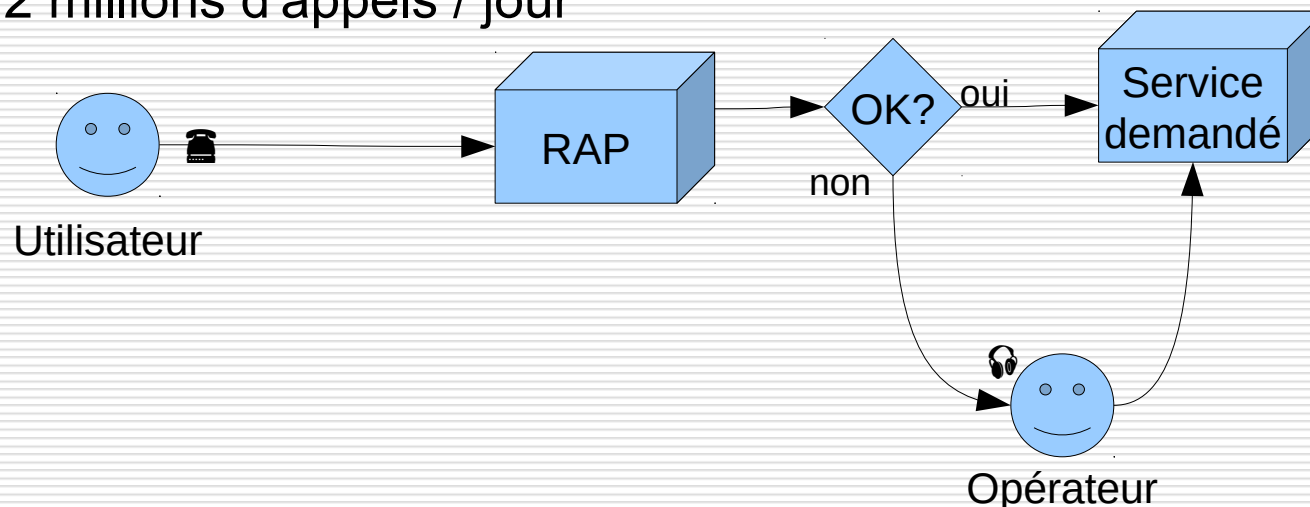
- Quels outils proposer pour aider ce dialogue ?

Domaines

- Informatique / TALN / Sciences cognitives

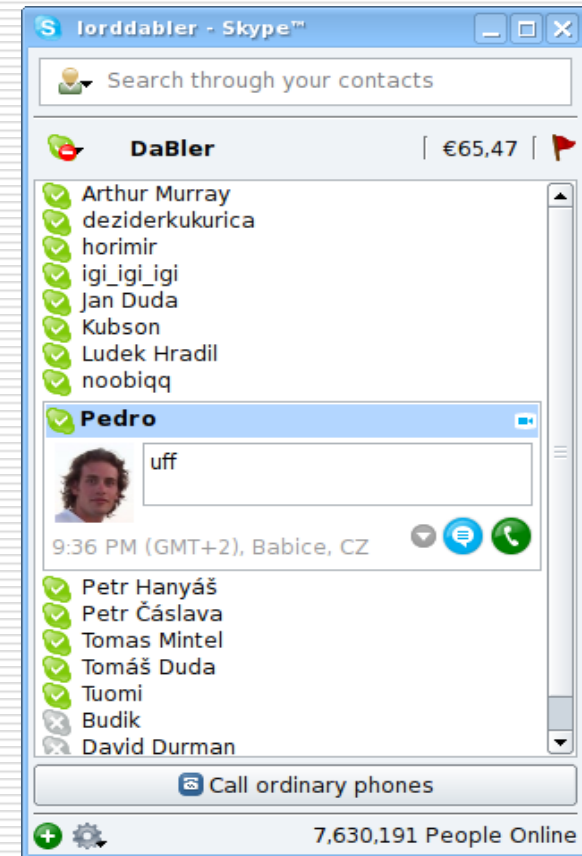
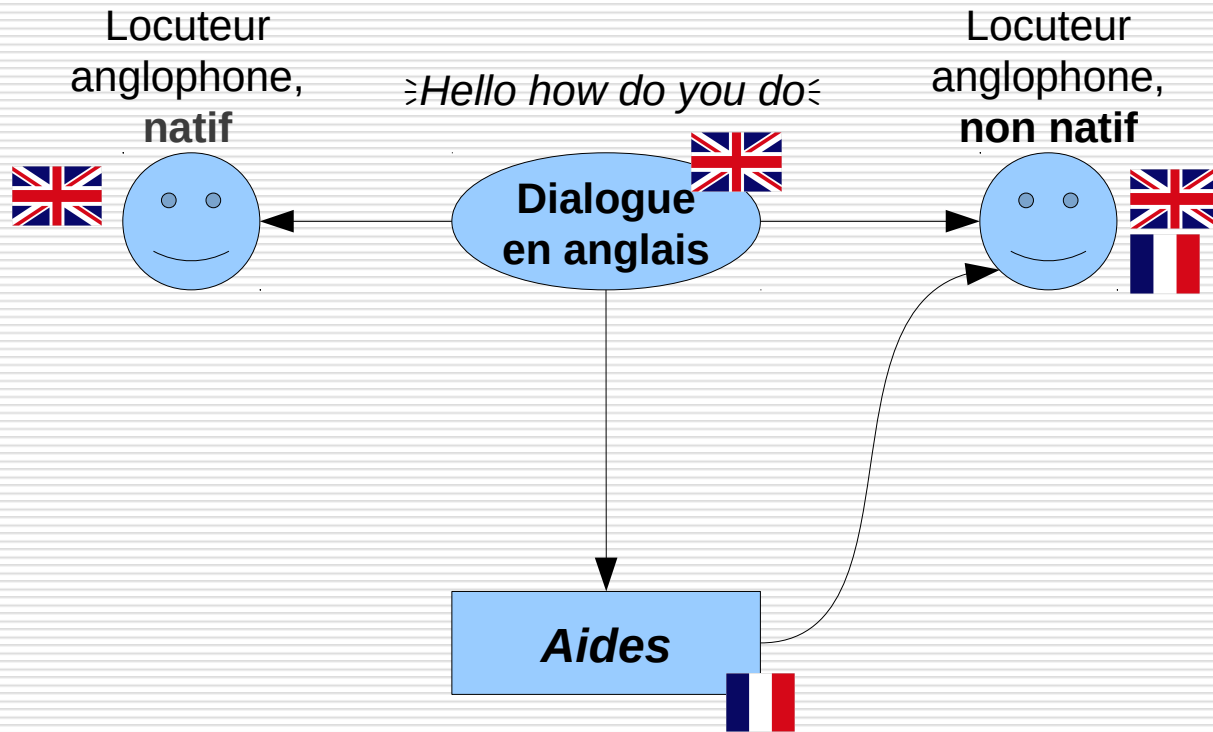
Contexte de la thèse

- Prosodie
 - Opérateur de services vocaux grands comptes (120 services)
 - Filiales en Espagne, aux États-Unis et au Canada
- Sensibilisé à la TA de parole en 2000
- Introduction de la RVAO en 2003
 - Reconnaissance Vocale Assistée par *Opérateur*
 - Aiguillage d'appels
 - 2 millions d'appels / jour



Nouvelle idée de service

Service sur *softphone* (lien avec la TA de parole)



- Thèse CIFRE (2005)
 - Spécification d'un produit
 - Développement et évaluation d'un prototype

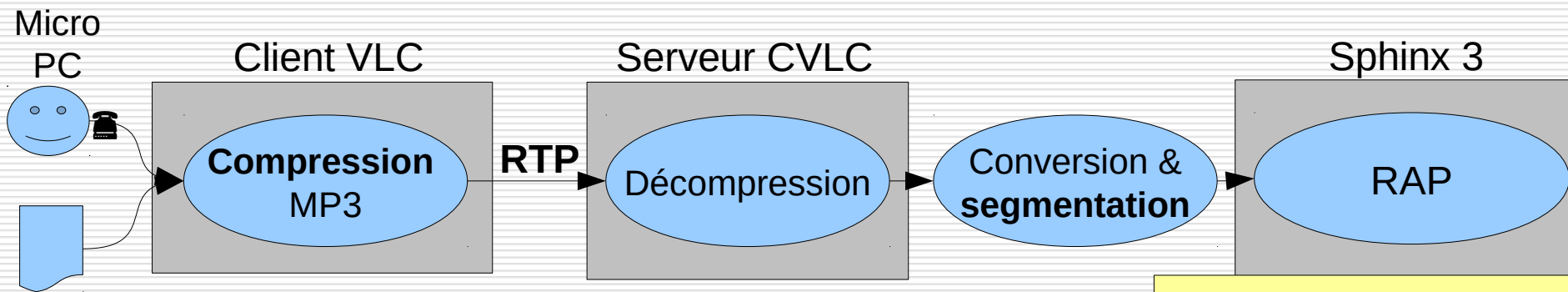
Première étude de faisabilité

- **Évaluation de la reconnaissance vocale**
 - Multilocuteur, domaine ouvert, parole non-native
 - État de l'art
 - 2003 : expérimentations sur domaine fermé [Wang & Schultz 03]
 - système Janus / données Verbmobil
 - WRR 85% (natif) → 65% (non-natif)
 - 2008 : thèse de Tan Tien Ping
 - Quelles performances en domaine ouvert ?

Expérimentation autour de la RAP

Test sur domaine ouvert

- Développement d'une plateforme expérimentale (10-11/2005)



Modèle de langage:
64K mots
(English Gigaword Corpus)

Modèle acoustique:
Parole continue, téléphone

- Validation de la plateforme sur parole native
 - WRR 60%
 - Temps de reco = 4 fois le temps de parole
(sur VIA C7 1,5 GHz, 2Go RAM)
- Test sur de la parole non-native (sources variées – 12/2005)
 - WRR < 1%

Redéfinition de l'objectif

Bilan de l'évaluation

- Reconnaissance vocale pas encore utilisable dans le cadre de ce travail
- Rappel de la contrainte CIFRE → un outil fonctionnel complet

Nouvel objectif (réaliste)

- Support du tchat multilingue

Plan

1. *La matière première à étudier*

- Tchat monolingue
- Tchat en langue seconde
- Dialogue en langue seconde

2. Aides informatisées nécessaires

3. Développement d'un prototype (*Koinè*)

4. Conclusion et perspectives

Plan

1. *La matière première à étudier*

- Tchat monolingue
- Tchat en langue seconde
- Dialogue en langue seconde

2. Aides informatisées nécessaires

3. Développement d'un prototype (*Koinè*)

4. Conclusion et perspectives

Qu'est-ce que le tchat ?

□ Historique

- Talk (1983) → IRC (1988) → ICQ (1996)

□ Principe de fonctionnement

- Client-serveur
- Communication écrite, dactylographiée
- Communication quasi-synchrone
 - Tout de même un temps dédié à la planification avant l'envoi du message

□ Organisation

- Canaux de discussion
 - Généralistes
 - Thématiques
- Tout est *message* : discussion, commandes, notifications du système

Phénomènes saillants du tchat (1/4)

□ Fautes d'orthographe

- [01] A> ptit question si il y pas le meme nombre d'element ds les 22 liste tu fais comment ?
- [02] A> car faut pas oublier que y auras pas que 2 conecction
- [03] B> A: tu dois tjs avoir le meme nombre d'element :)
- [04] A> faut
- [05] A> faux
- [06] A> enfin sauf lors d'une deco ou d'un connect
- [07] B> mais a par pdt ces brefs instants a priori tu as pas de prob
- [08] B> et puis au pire tu mets NULL dans le deuxieme membre de ta pere
- [09] A> imagine A B C D or A et connect avec B alors que B et connect avec A C et D donc pas le meme nombre d'élément
- [10] B> s/pere/paire/
- [11] B> hum il est bizarre ton rezo :)

Phénomènes saillants du tchat (2/4)

□ Abréviations

- [01] A> ptit question si il y pas le meme nombre d'element ds les 22 liste tu fais comment ?
- [02] A> car faut pas oublier que y auras pas que 2 conecction
- [03] B> A: tu dois tjs avoir le meme nombre d'element :)
- [04] A> faut
- [05] A> faux
- [06] A> enfin sauf lors d'une deco ou d'un connect
- [07] B> mais a par pdt ces brefs instants a priori tu as pas de prob
- [08] B> et puis au pire tu mets NULL dans le deuxieme membre de ta pere
- [09] A> imagine A B C D or A et connect avec B alors que B et connect avec A C et D donc pas le meme nombre d'élément
- [10] B> s/pere/paire/
- [11] B> hum il est bizarre ton rezo :)

Phénomènes saillants du tchat (3/4)

- Oralisation
- Marqueurs intonatifs

[01] A> petit question si il y pas le meme nombre d'element ds les 22 liste tu fais comment ?

[02] A> car faut pas oublier que y auras pas que 2 conecction

[03] B> A: tu dois tjs avoir le meme nombre d'element :)

[04] A> faut

[05] A> faux

[06] A> enfin sauf lors d'une deco ou d'un connect

[07] B> mais a par pdt ces brefs instants a priori tu as pas de prob

[08] B> et puis au pire tu mets NULL dans le deuxieme membre de ta pere

[09] A> imagine A B C D or A et connect avec B alors que B et connect avec A C et D donc pas le meme nombre d'élément

[10] B> s/pere/paire/

[11] B> hum il est bizarre ton rezo :)

Phénomènes saillants du tchat (4/4)

- Aides
- Corrections

[01] A> ptit question si il y pas le meme nombre d'element ds les 22 liste tu fais comment ?

[02] A> car faut pas oublier que y auras pas que 2 conecction

[03] B> A: tu dois tjs avoir le meme nombre d'element :)

[04] A> faut

[05] A> faux

[06] A> enfin sauf lors d'une deco ou d'un connect

[07] B> mais a par pdt ces brefs instants a priori tu as pas de prob

[08] B> et puis au pire tu mets NULL dans le deuxieme membre de ta pere

[09] A> imagine A B C D or A et connect avec B alors que B et connect avec A C et D donc pas le meme nombre d'élément

[10] B> s/pere/paire/

[11] B> hum il est bizarre ton rezo :)

Explications des ces phénomènes

- **Études existantes** [Pierozack 03]
 - Les utilisateurs
 - S'entraident
 - Sont conscients des problèmes
 - Font des *fautes* (volontaires ou non)
 - Se corrigent si nécessaire
 - Pourquoi ces *fautes* ?
 - Involontaires
 - Planification = production
 - Volontaires
 - Jeux avec la langue
 - Effet de communauté
 - Observations qualitatives faites sur ~37k mots utiles

Vers des données de grande taille ...

Exploitation et poursuite du travail de M2R

- **Situation en 2004 (M2R)**
 - Un seul corpus (italien 847 223 mots), format texte

- **Collecte pour le français (2004-2005)**
 - Logs du serveur EpikNet (pages HTML)
 - Questions d'éthique et de droit
 - Récupération des *logs*
 - Formatage du corpus vers XML
 - **Bilan : 105 canaux, 3 mois, 23 011 876 mots utiles**

Interface web (<http://aiakide.net/pagepro/corpuschat2/>)

Interface web (<http://aiakide.net/pagepro/corpustchat2/>)

[illegible]

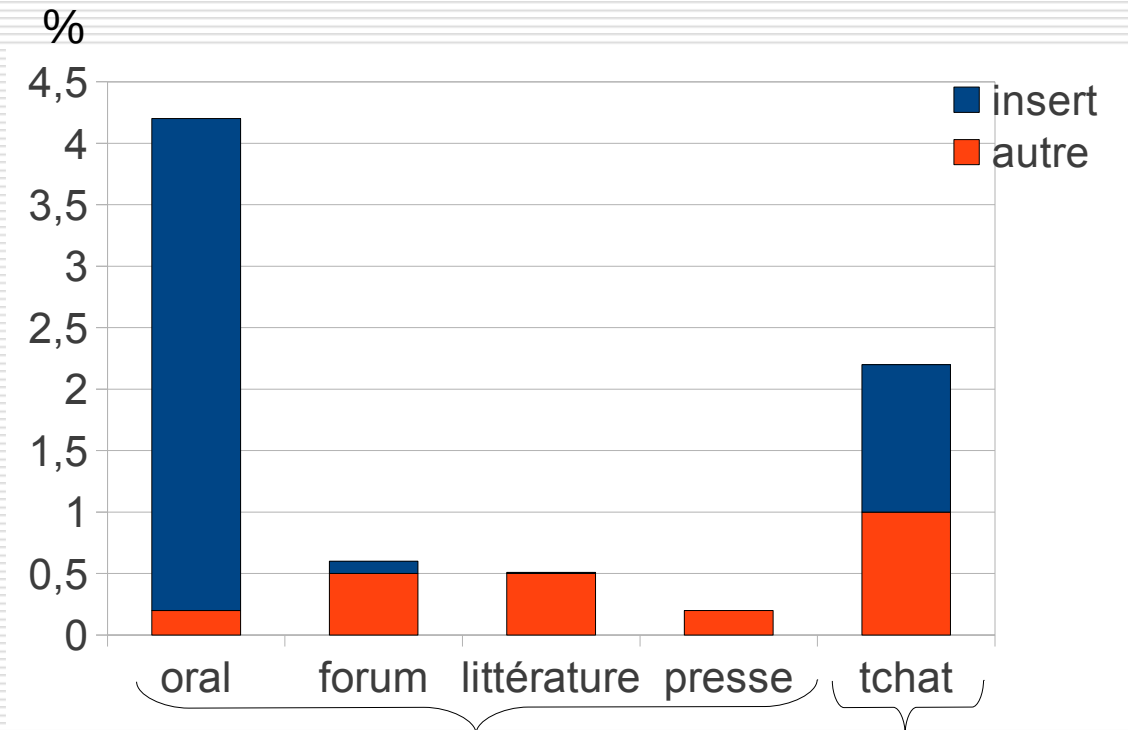
Interface web (<http://aiakide.net/pagepro/corpuschat2/>)

Interface web (<http://aiakide.net/pagepro/corpustchat2/>)

			oui	
596	7/1/2004	10:31	fred60-tonnton	Message grrrrrrrrrrrrrrr
597	7/1/2004	10:31	samo	Message Ben pourquoi t'es tout(e) grognon ?? Que t'arrive t - il fred60-tonnton ?
598	7/1/2004	10:31		Événement (windows@Ep.yP5bJrIEKg) vient d'arriver ! #edelweiss.
599	7/1/2004	10:31		Événement changement de mode(s) '+v jofazer' par Sphinx!services@oIC1HiORmEYDI
600	7/1/2004	10:31		Événement Changement de pseudo: fred60-tonnton > fred60
601	7/1/2004	10:31		Événement Changement de pseudo: fred60 > fred60-tonnton
602	7/1/2004	10:31	guetty	Message kssssssssssssssssssssssssssssss jofazer
jofazer	Message	comment va ma guetty ce amtin?		
guetty	Message	trankil jofazer et toi ?		
	Événement	(AS230B@EplHHItu4RUNc) vient d'arriver ! #edelweiss.		
boss1	Message	bonjour tt l'monde		
jofazer	Message	moi pas trop mais bon, pas grave		
jofazer	Message	bonjour boss1		
boss1	Commande	!bisous KaRiNe,guetty,Adeline,HELENE33		
samo	Message	~~* boss1 *~~ fait un gros bisou à ~~ KaRiNe,guetty,Adeline,HELENE33 *~~ Smouack ! d:-(p ()-:O		
617	7/1/2004	10:33	guetty	Message kssssssssssssssssssssss boss1
618	7/1/2004	10:34	HELENE33	Message bisouss boss1
619	7/1/2004	10:34		Événement (PrinCeZZe@chqY6admLxSa2) s'est déconnecté: Quit: Ne me demandez pas pourquoi je l'aime.....je devrais vous expliquer pourquoi ie vis

Résultats obtenus (1/2)

- Confirmation des constats de [Piérozack 03]
- Comparaison avec d'autres genres



Proportion d'occurrences de *Bon!* comme insert.

Un excellent marqueur de l'oralité selon [Benzitoun & al. 04].

Étude équipe DÉLIC
144k mots
[Benzitoun & al. 04]

Mes résultats
35k mots

→ le tchat est la forme d'écrit qui se rapproche le plus de l'oral

Résultats obtenus (2/2)

□ Comparaison avec les SMS

■ Plus normalisé

Formes	SMS	Tchat
ptet	663	29
ptetre	133	8
p-e	923	6
pe etre	0	3
pe	600	0,4
ptetr	223	0,3

Nombre d'occurrences
pour 100k mots

SMS: échantillon de 30k
messages

Corpus *SMS pour la
science* [Fairon 06]

Tchat: échantillon de 1849k
messages de mon corpus

- SMS: **p-e** > **ptet** > **pe** >> ptetr > ptetre
- Tchat: **ptet** >> ptetre > p-e > pe etre

Premiers éléments d'un cahier des charges

Issus de l'étude du tchat monolingue

- **Fonctionnalités à ajouter**
 - Correction de messages *a posteriori*
 - Les utilisateurs corrigent leurs messages

- **Fonctionnalités à préserver**
 - Pouvoir utiliser de l'écrit non-standard
 - Les utilisateurs enrichissent l'écrit pour véhiculer du phatique/paraverbal

Plan

1. La matière à étudier

- Tchat monolingue
- Tchat en langue seconde
- Dialogue en langue seconde

2. Propositions et spécifications

3. Architecture et implémentation

4. Conclusion et perspectives

Qu'est-ce que le tchat L2 ?

□ **Présentation**

- Tchat avec sous-titrage automatique

□ **Études existantes**

- Didactique des langues [Sotillo 05-06]
- Pas de différences majeures avec le dialogue L2 oral

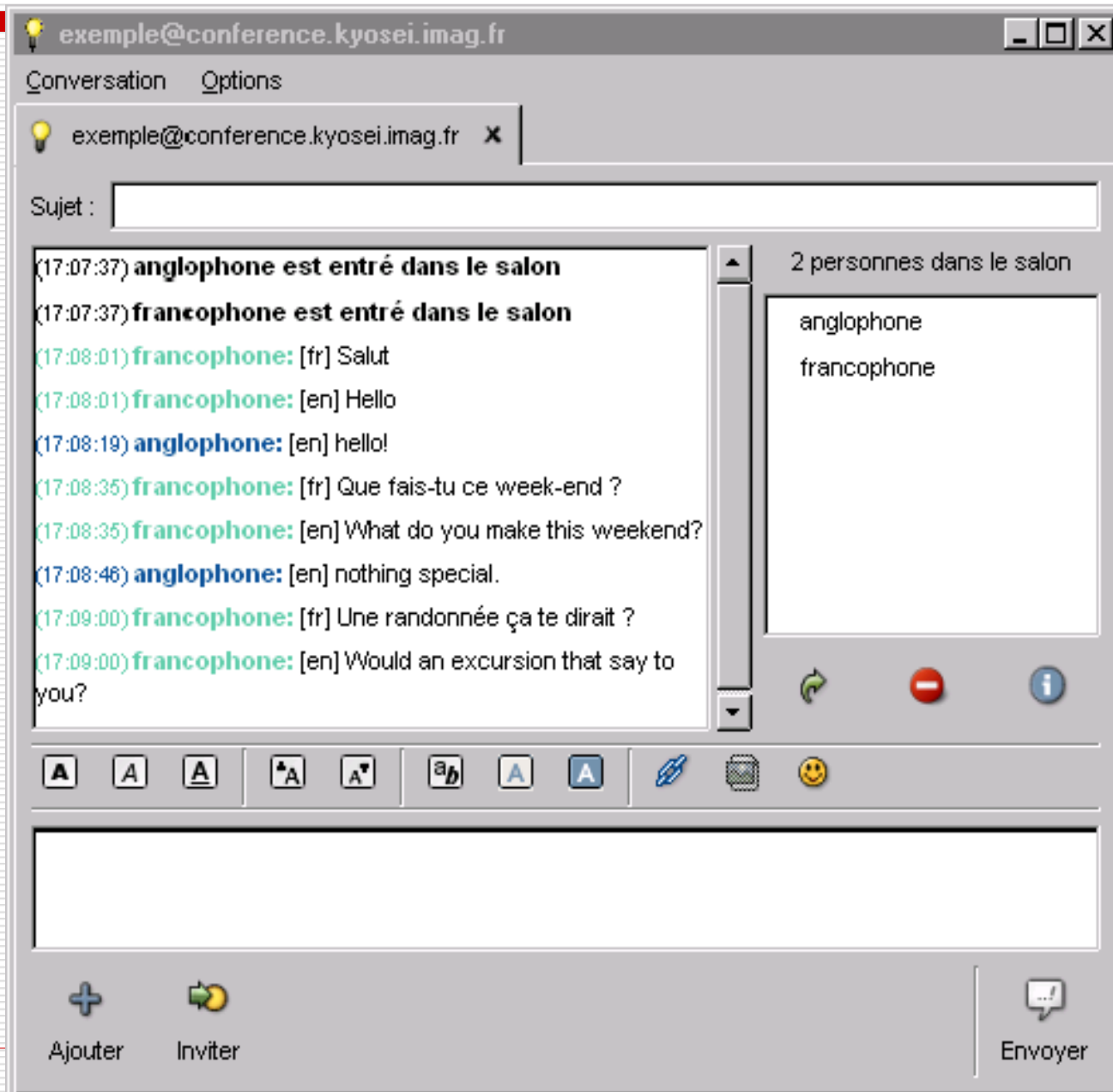
□ **Systèmes existants**

- *Wordlingo Chat, AmiChat, Language Grid Playground, ChatSphere, Qopuchawi, etc.*

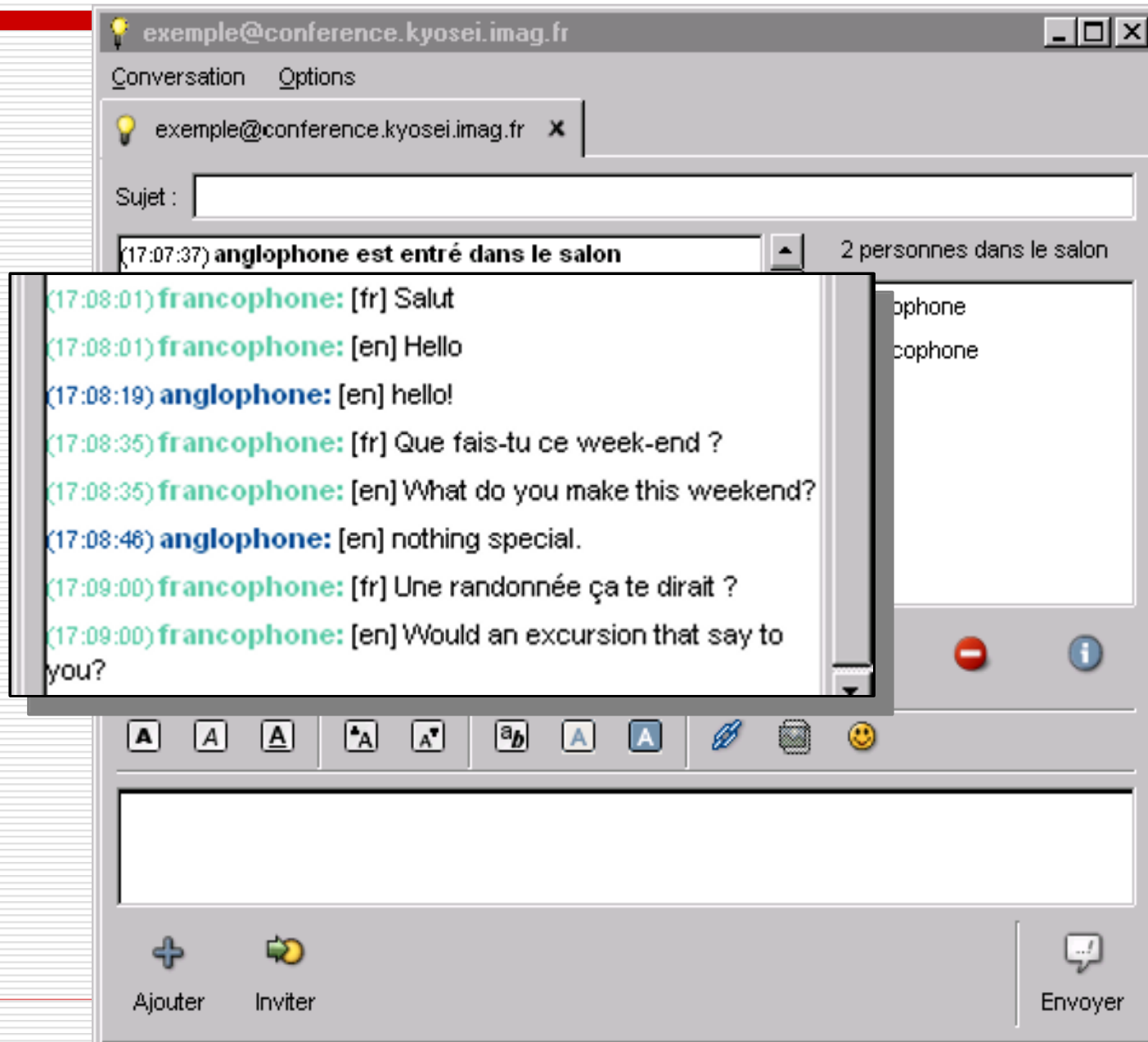
□ **Expérimentation en M2R**

- Développement d'une passerelle de tchat multilingue

Expérimentation sur le tchat L2



Expérimentation sur le tchat L2



Nouveaux éléments d'un cahier des charges

Issus de l'étude du tchat en langue seconde

□ Limites des outils existants

- Pas de correction des messages *a posteriori*
- Pas de contrôle de la TA
 - Ni *a priori* (validation)
 - Ni *a posteriori* (postédition/correction)

□ Approfondissement de l'étude

- Très peu de données sur le **tchat** en langue seconde
- De nombreuses études sur le dialogue en langue seconde **oral**

Plan

1. *La matière première à étudier*

- Tchat monolingue
- Tchat en langue seconde
- Dialogue en langue seconde

2. Aides informatisées nécessaires

3. Développement d'un prototype (*Koinè*)

4. Conclusion et perspectives

Le dialogue en langue seconde

□ Principales écoles de pensée

- Didactique des langues
- Anglais langue internationale
 - Grec *koinè*, latin médiéval, *global English*

□ Mon approche

- Suivre l'école *Anglais international*
- Mettre en œuvre des outils existants de traitement de la langue standard

Phénomènes saillants du dialogue L2

Entraide

- Corrections

Phénomènes saillants du dialogue L2

Entraide

- Corrections

NS1: and we er: developed (.) (to) (.) store fit
er size rack. er which er (.) i er: showed the

NS4: mhm

NS1: pictures on page twenty-eight (.) and
also we called it wire rack (.) so that was a
bi- er: that was our er [A] and [P] er

NNS2: *seventeen*

NS1: page seventeen?

NNS2: mhm

NS1: oh yeah pa- sorry page seventeen

NNS5: mhm (.)

NS1: we produced about three hundred of
that

NS4: *of those?*

NNS2: yeah

Phénomènes saillants du dialogue 12

Entraide

- Corrections
- Sous-dialogues

NS: what are you mumbling?

NNS: that song hhhhm. hh. **what's that woman's name? that singer... the *chocolate woman*!**

NS: areta franklin

NNS: no:::!! she's not black!

NS: you said she was black!

NNS: no I did not!

NS: yes you did!

NNS: NO I DIDN'T

NS: yes you did. you said that chocolate woman.

NNS: yes, the one that sings that your love is better than chocolate

NS: sarah maclofn? she's white

NNS: I never said she was black!

NS: yes (.) you (.) did

NNS: NO I didn't ((frustrated))

NS: ((laughs))

NNS: whats so funny↑ ((quite angry))

NS: honey, **a *chocolate woman* IS a black woman.**

NNS: oh... how am I supposed to know that?

Synthèse des besoins des utilisateurs

Importance du phatique/paraverbal

- Ne pas faire perdre la face
 - On doit faire passer plus de phatique
- Mettre les formes... mais on ne sait pas comment !
 - On passe encore plus par du paraverbal/extra-linguistique

Problèmes des locuteurs

- Manque de vocabulaire
- Manque de connaissances grammaticales
 - Construction syntaxique
 - Usage

Nouveaux éléments d'un cahier des charges

Issus de l'étude du dialogue en langue seconde

Importance du phatique/paraverbal

- Tolérance au paraverbal (comme en tchat)

Quels sont les problèmes des locuteurs ?

- Vocabulaire → dictionnaire
- Manque de connaissances grammaticales
 - Construction syntaxique → traduction automatique
 - Usage → concordancier

Plan

1. *La matière première à étudier*
2. Aides informatisées nécessaires
3. Développement d'un prototype (*Koinè*)
4. Conclusion et perspectives

Cahier des charges

- ❑ **Correction de messages**
- ❑ **Traiter de l'écrit non standard**
- ❑ **Aides linguistiques**
 - Dictionnaire
 - Traduction automatique
 - Concordancier
- ❑ **Contrôle/correction des aides**
- ❑ **Faciliter les sous-dialogues méta-linguistiques**

Écrit non standard : angles d'approches

□ **Approches possibles**

- Développer des outils et ressources spécifiques pour le tchat
 - Dictionnaires bilingues, traduction automatique, etc.
- Adapter le tchat vers de l'écrit standard
 - Réutilisation des outils et ressources existantes

□ **Approche retenue**

- Adaptation vers de l'écrit standard
 - Démarche similaire à celle adoptée par la communauté de la traduction de parole

Écrit non standard : proposition

□ 3 classes de *mots*

- Mots formels (mots correctement orthographiés)
- Mots formels *oralisés*, et donc formalisables :
 - À l'aide d'un dictionnaire (abréviations, formes euphoniques)
 - À l'aide d'heuristiques (allongements)
- Mots oraux (émoticons), repérables :
 - À l'aide d'un dictionnaire

□ Que faire des mots inconnus ?

- Les utilisateurs sont conscients des problèmes, et capables d'effectuer des corrections
- On peut signaler le problème

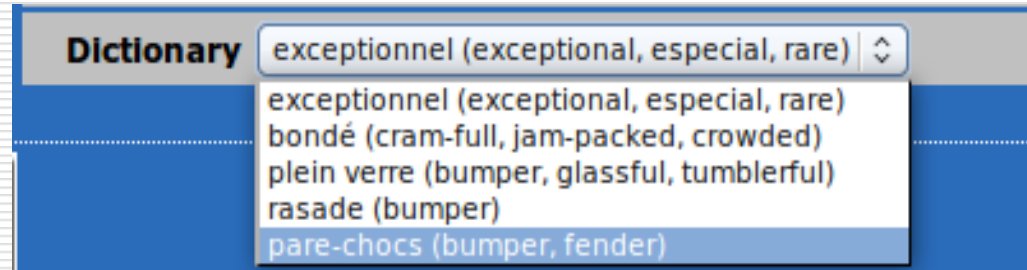
Aides linguistiques

anglophone

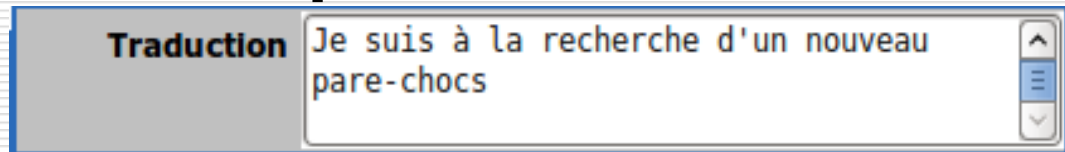
14:49:14

I am looking for a new bumper

□ Dictionnaire



□ Traduction automatique



□ Concordancier

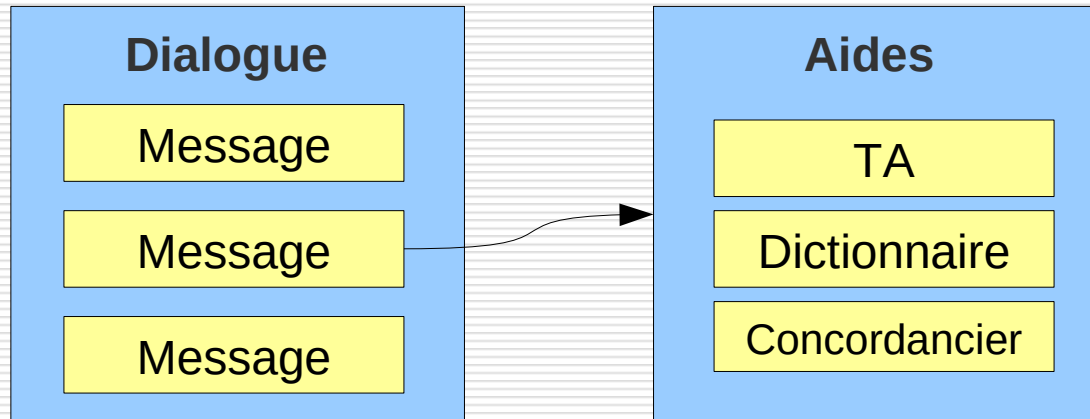
26700 occurrences de ce terme

I want you to come here (ONE TIME) Please sir release her HEY!

I Want You to Know That There Is Always Hope is the debut and only album

Aides linguistiques

- Organisation des aides
 - Proposition d'une approche à médiation «faible»



My language: English

francophone
14:52:31

Hello how do you do :-)

Hello how do you do :-)

anglophone
14:53:08

Looking for a new grille

Looking for a new grille

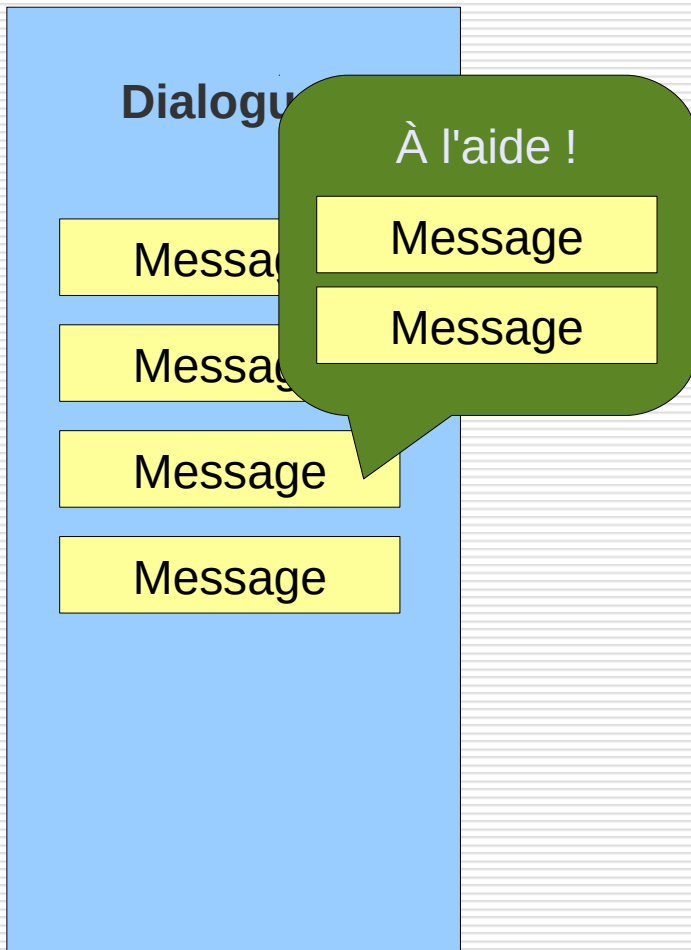
Comments francophone: what do you mean ?
Close

Translation
la recherche d'une nouvelle grille

Dictionary calandre (roller, calender, radiator grill)

Send

Faciliter les sous-dialogues méta-linguistiques



anglophone
16:56:20

Is it a worked over bumper ?

Is it a **worked over** bumper ?

Comments francophone: What do you mean ?
anglophone: not a new one
francophone: OK I see

Close

Plan

1. *La matière première à étudier*
2. Aides informatisées nécessaires
3. Développement d'un prototype (*Koinè*)
4. Conclusion et perspectives

Vers une évaluation de l'approche

Objectif: évaluation en ligne

- Évaluation en situation réelle
- Outils de dépannage : l'utilisateur ne doit rien avoir à installer...
 - ... mais simplement une URL à donner
- Développement sous forme d'application web : *Koinè*

Cahier des charges d'une bibliothèque de support

Montée en complexité du développement

- Site web statique (HTML)
 - Collection de pages
- Site web dynamique (PHP, J2EE)
 - + contenu des pages dépend de l'utilisateur
 - + formulaires
- Application web (AJAX)
 - Application *traditionnelle*, mais client-serveur

Des problèmes génériques

- Décharger le développeur de ces problèmes
- Développement d'une bibliothèque indépendante : *Krater*

Cahier des charges d'une bibliothèque de support

Montée en complexité du développement

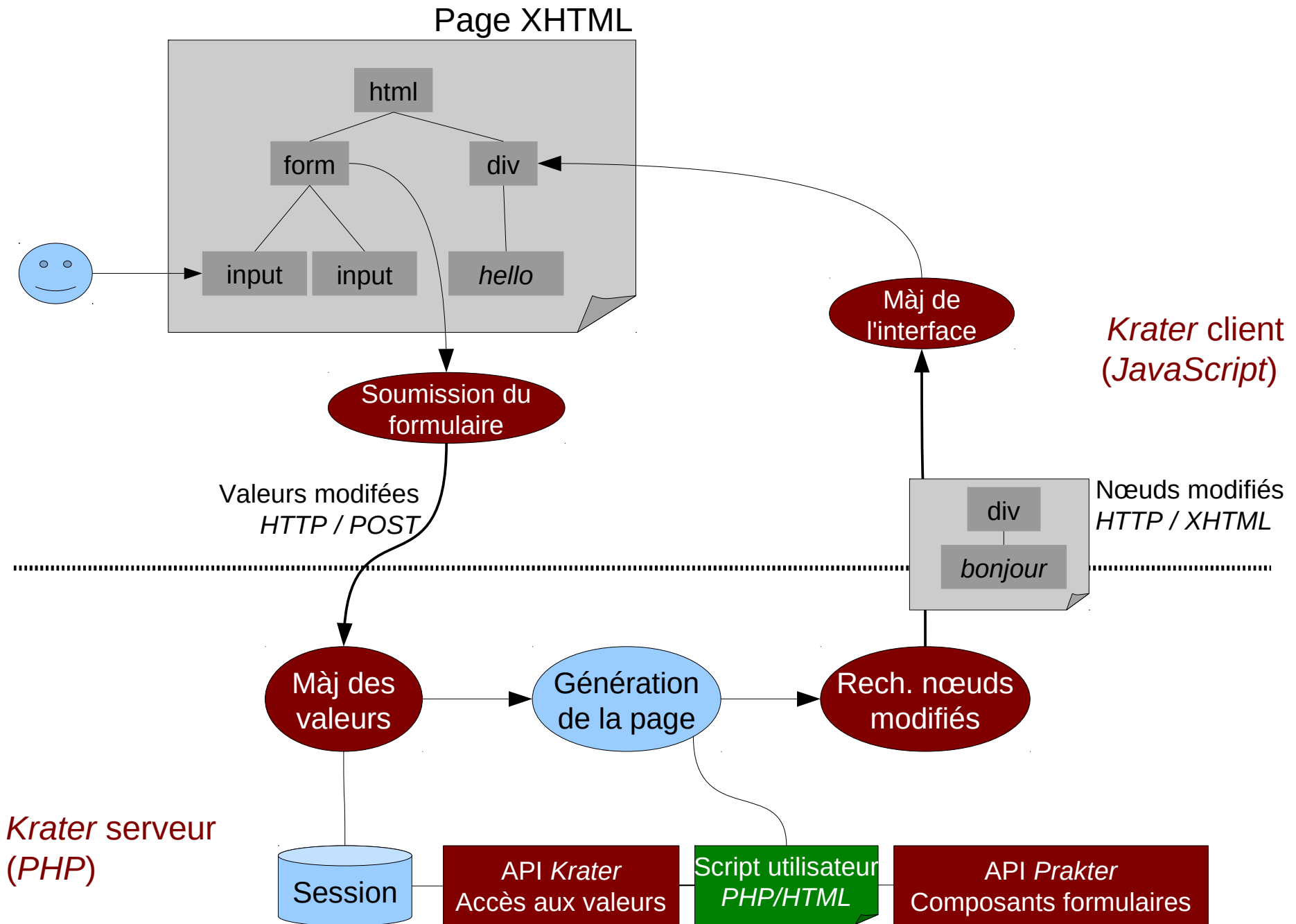
- Site web statique (HTML)
 - Collection de pages
- Site web dynamique (PHP, J2EE)
 - + contenu des pages dépend de l'utilisateur
 - + formulaires
- Application web (AJAX)
 - Application *traditionnelle*, mais client-serveur

Des problèmes génériques

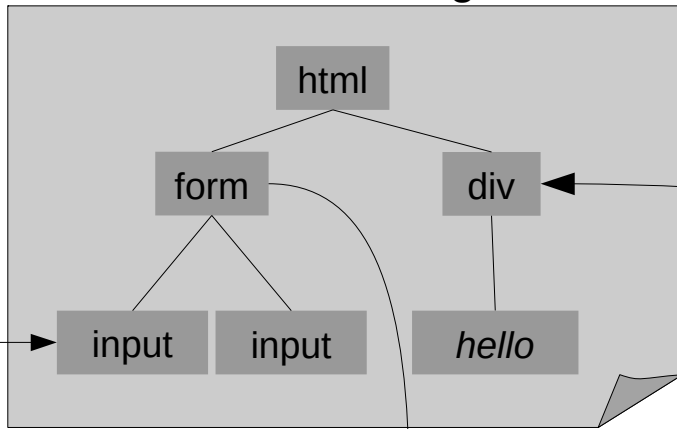
- Décharger le développeur de ces problèmes
- Développement d'une bibliothèque indépendante : *Krater*

Développement de la bibliothèque *Krater*

- **Fin 2006 : effervescence autour des applications web**
- **Première approche : utilisation de bibliothèques existantes**
 - 2007: Firefox / XPFE (Mozilla)
 - Début 2008: XFORMS (W3C)
- **Seconde approche: réimplémentation (fin 2008)**
 - À partir de *fondamentaux* éprouvés (PHP, JavaScript, XML)
- **Évaluation**
 - Utilisable pour du prototypage ou le développement d'interfaces lourdes
 - Générique et réutilisable
 - Projet CIFLI-SurviTra (LIG/IIT) <http://aiakide.net/survitra-3/>
 - Projet Scientext (LIDILEM) <http://aiakide.net/scientext10/>



Page XHTML



Soumission du formulaire

Valeurs modifiées
HTTP / POST

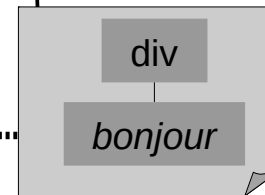
Màj des valeurs

Génération de la page

Rech. nœuds modifiés

Màj de l'interface

*Krater client
(JavaScript)*



Nœuds modifiés
HTTP / XHTML



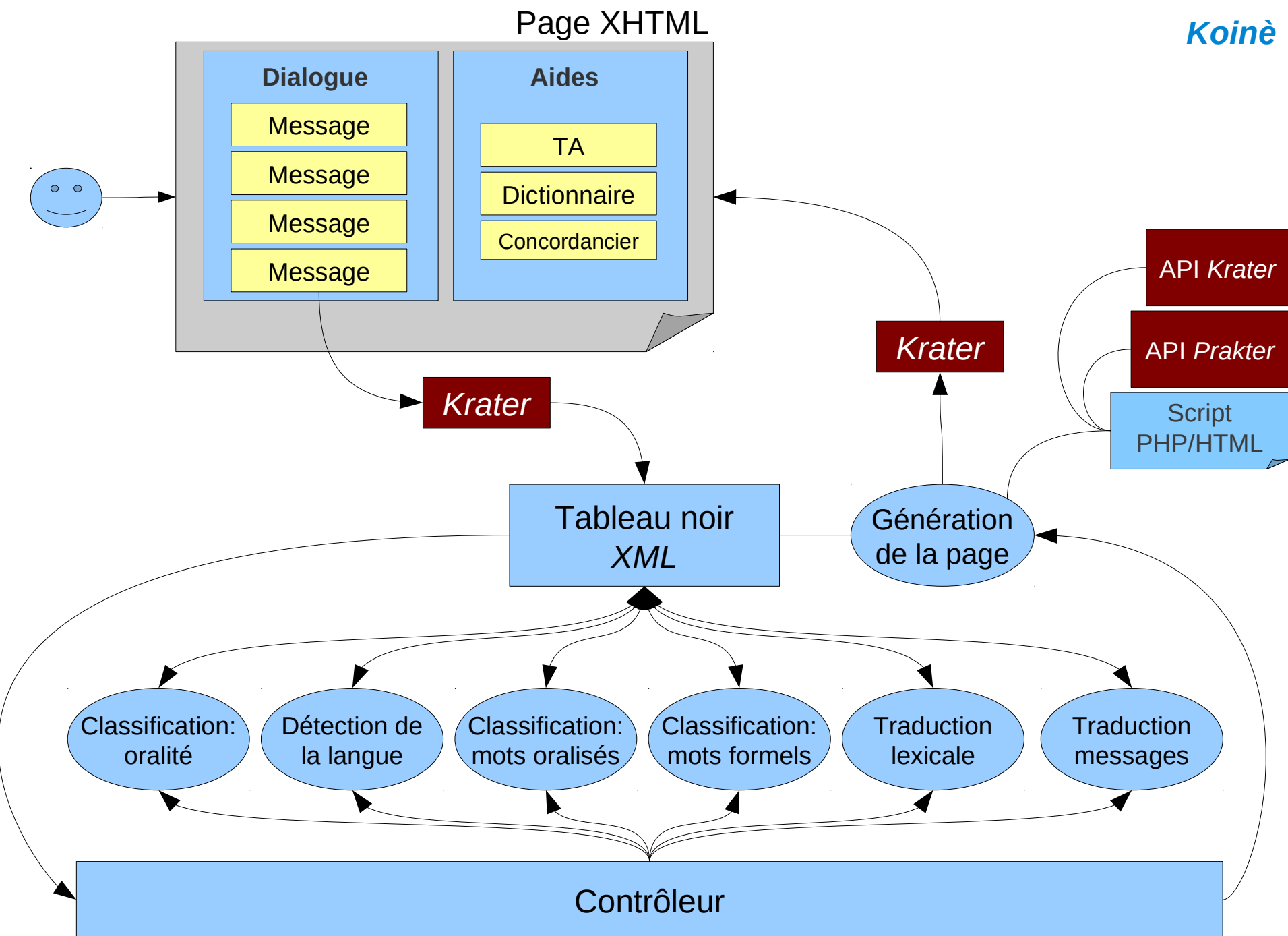
Session

API Krater
Accès aux valeurs

Koinè

API Prakter
Composants formulaires

*Krater serveur
(PHP)*



Démonstration

Plan

1. *La matière première à étudier*
2. Aides informatisées nécessaires
3. Développement d'un prototype (*Koinè*)
4. Conclusion et perspectives

Apports (1/3)

Étude du tchat (linguistique)

- Corpus (24M mots)
 - Outil de navigation et de recherche
- Confirmation de résultats antérieurs
- Mise en perspective par rapport à d'autres modalités

Aide au tchat monolingue (TALN)

- Correction de messages
 - Prise en charge de l'écrit non-standard
-

Apports (2/3)

Aide au dialogue en langue seconde (TALN)

- Mise en évidence des problèmes
- Propositions motivées d'aide au dialogue L2
 - Aides linguistiques
 - Dictionnaire
 - Traduction automatique
 - Concordancier
- Contrôle/correction de ces aides
- Faciliter la coopération méta-linguistiques
 - Via un sous-dialogue
 - Via les aides

Apports (3/3)

Plateforme (informatique et TALN)

- Pour le prototypage d'applications web (*Krater*)
- Pour le déploiement d'aides au tchat (*Koinè*)
 - Implémentation d'une partie des aides

Perspectives

- **Vers un prototype complet ...**
 - Implémenter quelques fonctionnalités manquantes
 - Optimiser quelques algorithmes

- **... permettant de conduire des évaluations**
 - Quantitatives
 - En laboratoire
 - Mesurer le gain de productivité
 - Qualitatives
 - Sur le web
 - Évaluer l'apport de chaque type d'aide, l'approfondissement des conversations, etc.

Publications et projets

Publications

- 2009 G. Fafiotte, A. Falaise, J. Goulian, «**CIFLI-SurviTra, deux facettes: démonstrateur de composants de TA fondée sur UNL, et phrasebook multilingue**», Démonstration, Actes TALN 2009, Senlis.
- 2007 C. Boitet, P. Bhattacharyya, E. Blanc, S. Meena, S. Boudhh, G. Fafiotte, A. Falaise, V. Vacchani, «**Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction**», actes de SNLP 2007, Pattaya (Thaïlande).
- 2005 A. Falaise, «**Constitution d'un corpus de français tchaté**», actes de RECITAL 2005, Dourdan.
- 2003 M. Faraco, M.-L. Barbier, A. Falaise, S. Branca Rosoff, «**Codage et traitement automatique de corpus pour l'étude de prises de notes en français langue première et langue seconde**», Arob@se 7.

Affiches et posters

- 2006 A. Falaise, **Le projet Tradphone: un softphone au secours des locuteurs non-natifs aux prises avec des difficultés de communication en langue étrangère**. Affiche pour l'école d'été Dialogue et Interaction, Autrans.
- 2006 A. Falaise, **Le traitement de la langue au service de la téléconversation en langue étrangère**. Affiche pour l'école de printemps Traitement des Connaissances, Apprentissage et NTIC, Évian.

Rapports et mémoires

- 2004 A. Falaise, **Premier pas vers une TA interactive pour le tchat**, rapport de stage de M2R, INPG.
- 2002 A. Falaise, **Traitement et analyse automatique d'un corpus de parole en vue de la génération de règles prosodiques pour la synthèse des nombres**, rapport de stage de maîtrise, université Stendhal.

Projets

- 2007-2009 **CIFLI-SurviTra** (LIG - Laboratoire d'Informatique de Grenoble / IIT - Indian Institute of Technology)
- 2008-2009 **Scientext** (LIDILEM - Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles)