



Achille Falaise
M2R ICA – option Sciences cognitives

Premier pas vers une TA interactive pour le tchat

21 juin 2004

Sous la direction de :

Christian Boitet

Hervé Blanchon

GETA-CLIPS

Jury :

Valérie Bellynck

Hervé Blanchon

Christian Boitet

Jeanny Hérault

Introduction

- ▶ Contexte
 - ▶ linguistique de corpus
 - ▶ TALN, traduction automatique (TA)
- ▶ Objectif
 - ▶ améliorer la TA de tchat
- ▶ Problèmes
 - ▶ quelles sont les caractéristiques de la langue « tchatée » ?
 - ▶ quels problèmes pose la TA « classique » de tchat ?
 - ▶ quelles améliorations apporter ?

Plan

1. Le tchat

1. qu'est-ce que c'est
2. un corpus de tchat
3. la langue tchatée

2. Le tchat multilingue

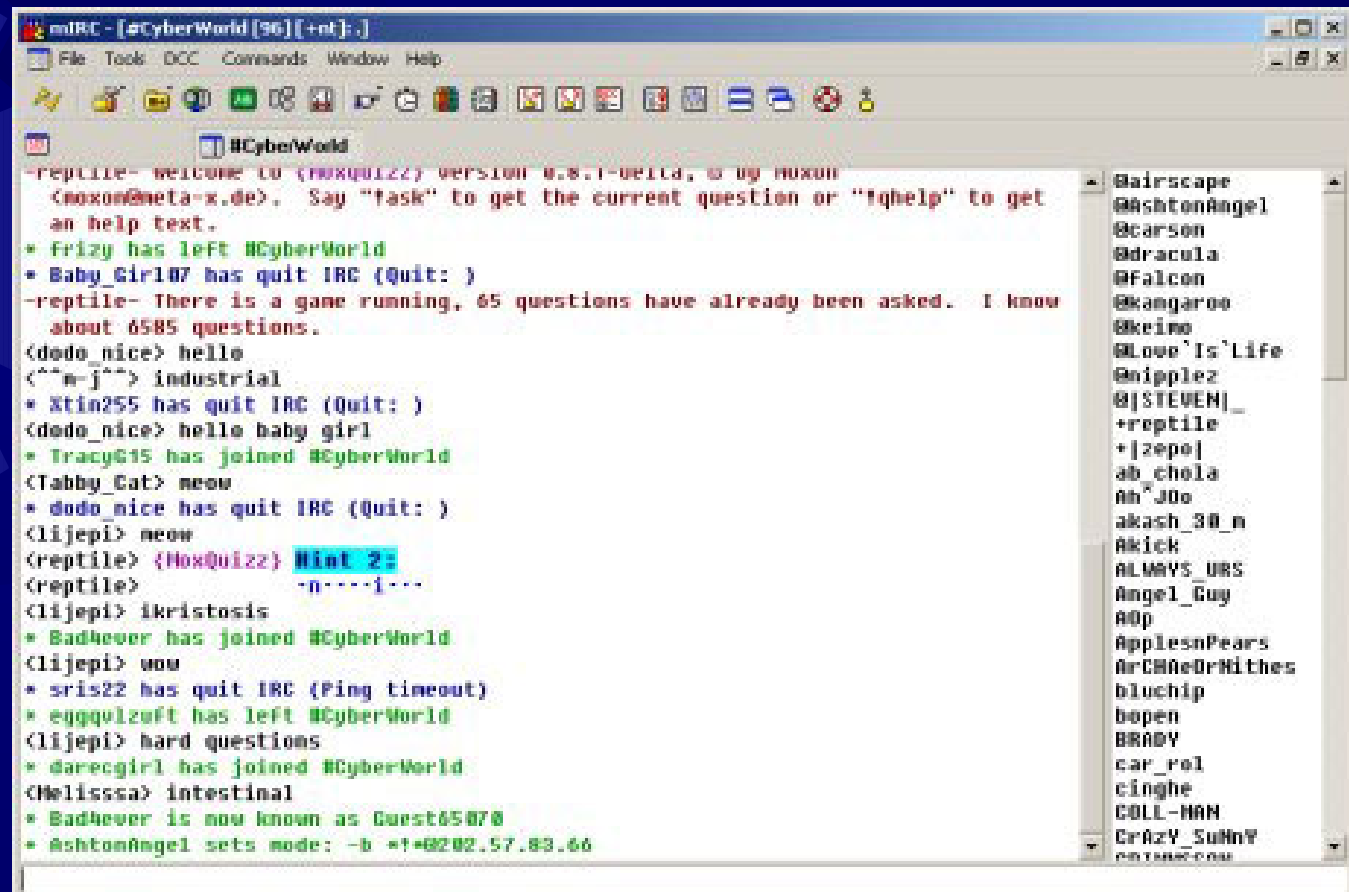
1. qu'est-ce que c'est
2. l'existant
3. collecte d'un corpus

Le tchat, qu'est-ce que c'est ?

- ▶ Outil de dialogue en ligne
 - ▶ en temps réel ($\neq$ e-mails, forums)
 - ▶ écrit ($\neq$ vidéoconférence)
- ▶ Permet de mener des conversations en ligne
 - ▶ au sein de canaux
- ▶ La messagerie instantanée
 - ▶ réseaux les plus connus : MSN, ICQ, AIM, etc...
 - ▶ en 2003 : 18% des internautes français
 - ▶ conversation limitée à deux utilisateurs à la fois
 - ▶ une sous-classe de tchat

Le tchat, qu'est-ce que c'est ?

- Du point de vue de l'utilisateur



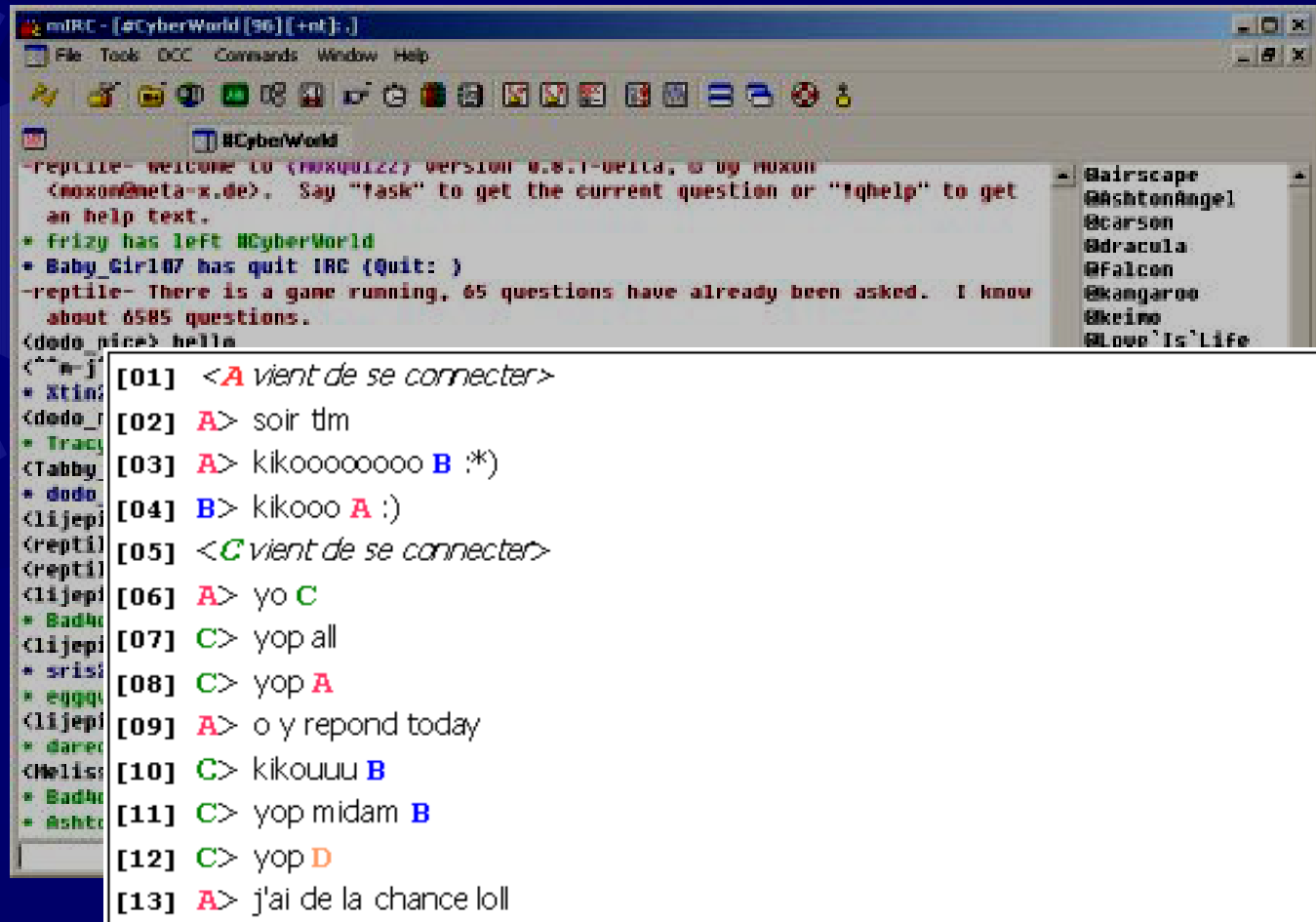
The screenshot shows the mIRC chat client interface. The main window displays a chat log with various messages, including welcome messages, user joins, and game-related text. The right-hand pane shows a list of users currently in the channel, including @Bairscape, @AshtonAngel, @carson, @dracula, @falcon, @kangaroo, @keino, @Love'Is'Life, @nipplez, @|STEVEN|, +reptile, +|zepo|, ab chola, Ah'J0o, akash_30_n, Akick, ALWAYS_URS, Angel_Guy, ADp, ApplesnPears, ArCHArDrMithes, bluchip, bopen, BRADY, car_ral, cinghe, COLL-NAN, CRAZY_SuNnY, and on time con.

```
mIRC - [#CyberWorld (36)] [+nt]:  
File Tools DCC Commands Window Help  
#CyberWorld  
-reptile- welcome to (moxquizz) version 0.8.1-beta, © by moxon  
(moxon@meta-x.de). Say "task" to get the current question or "tqhelp" to get  
an help text.  
* frizy has left #CyberWorld  
* Baby_Girl07 has quit IRC (Quit: )  
-reptile- There is a game running, 65 questions have already been asked. I know  
about 6585 questions.  
<dodo_nice> hello  
<^m-j^> industrial  
* Xtin255 has quit IRC (Quit: )  
<dodo_nice> hello baby girl  
* TracyG15 has joined #CyberWorld  
<Tabby_Cat> meow  
* dodo_nice has quit IRC (Quit: )  
<lijepl> meow  
<reptile> (MoxQuizz) Hint 2:  
<reptile> -n----i---  
<lijepl> ikristosis  
* Bad4ever has joined #CyberWorld  
<lijepl> wow  
* sris22 has quit IRC (Ping timeout)  
* eggquizzuft has left #CyberWorld  
<lijepl> hard questions  
* darecgirl has joined #CyberWorld  
<Melisssa> intestinal  
* Bad4ever is now known as Guest65878  
* AshtonAngel sets mode: -b +!+@202.57.83.66
```

Vue d'un client de tchat (mIRC)

Le tchat, qu'est-ce que c'est ?

- ▶ Du point de vue de l'utilisateur



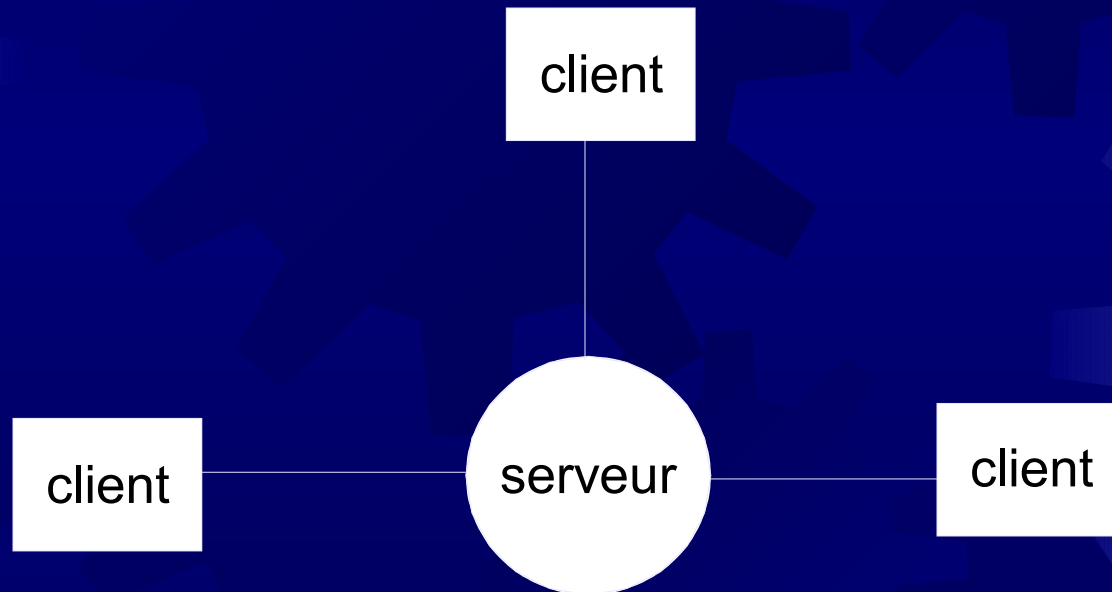
The screenshot shows an IRC window with a menu bar (File, Tools, DCC, Commands, Window, Help) and a toolbar. The main text area displays a chat log with messages from 'reptile-' and 'dodo_nice'. A list of users is visible on the right side of the window. A white box is overlaid on the bottom of the window, containing a list of chat messages numbered [01] to [13].

```
File Tools DCC Commands Window Help
#CyberWorld
-reptile- welcome to (moxquizz) version 0.8.1-delta, © by moxon
(moxon@meta-x.de). Say "task" to get the current question or "tqhelp" to get
an help text.
* Frizy has left #CyberWorld
* Baby_Girl07 has quit IRC (Quit: )
-reptile- There is a game running, 65 questions have already been asked. I know
about 6585 questions.
<dodo_nice> hello
<""-j
* Xtin:
<dodo_n
* Tracy
<Tabby_
* dodo
<lijepi
<reptil
<reptil
<lijepi
* Bad4u
<lijepi
* srisi
* eggqu
<lijepi
* damed
<Helis:
* Bad4u
* Ashto
```

[01] <A vient de se connecter>
[02] A> soir tm
[03] A> kikooooooooo B ;*)
[04] B> kikooo A :)
[05] <C vient de se connecter>
[06] A> yo C
[07] C> yop all
[08] C> yop A
[09] A> o y repond today
[10] C> kikouuu B
[11] C> yop midam B
[12] C> yop D
[13] A> j'ai de la chance loll

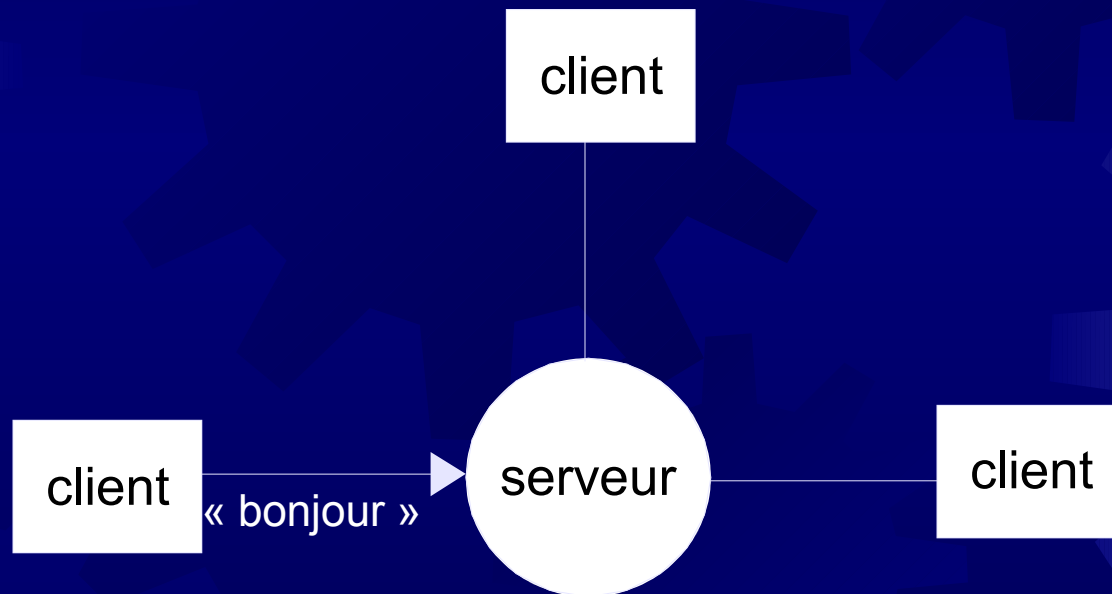
Le tchat, qu'est-ce que c'est ?

- ▶ D'un point de vue technique
 - ▶ architecture client-serveur



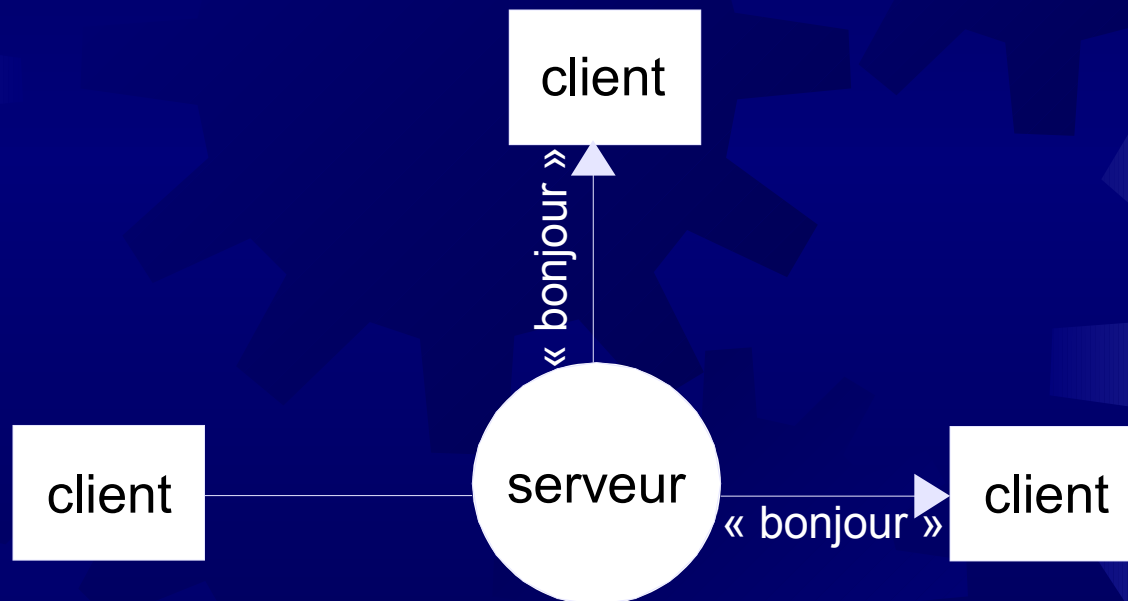
Le tchat, qu'est-ce que c'est ?

- ▶ D'un point de vue technique
 - ▶ architecture client-serveur



Le tchat, qu'est-ce que c'est ?

- ▶ D'un point de vue technique
 - ▶ architecture client-serveur



Un corpus de tchat

- ▶ Nécessité de constituer un corpus de tchat
 - ▶ voir quel type de langue on a à traiter
 - ▶ pas de grand corpus de tchat disponible
- ▶ Le corpus de départ
 - ▶ 18 millions de mots
 - ▶ obtenu à partir de logs de tchat publiés sur Internet

Un corpus de tchat

- ▶ Structure du corpus
 - ▶ format XML

Un corpus de tchat

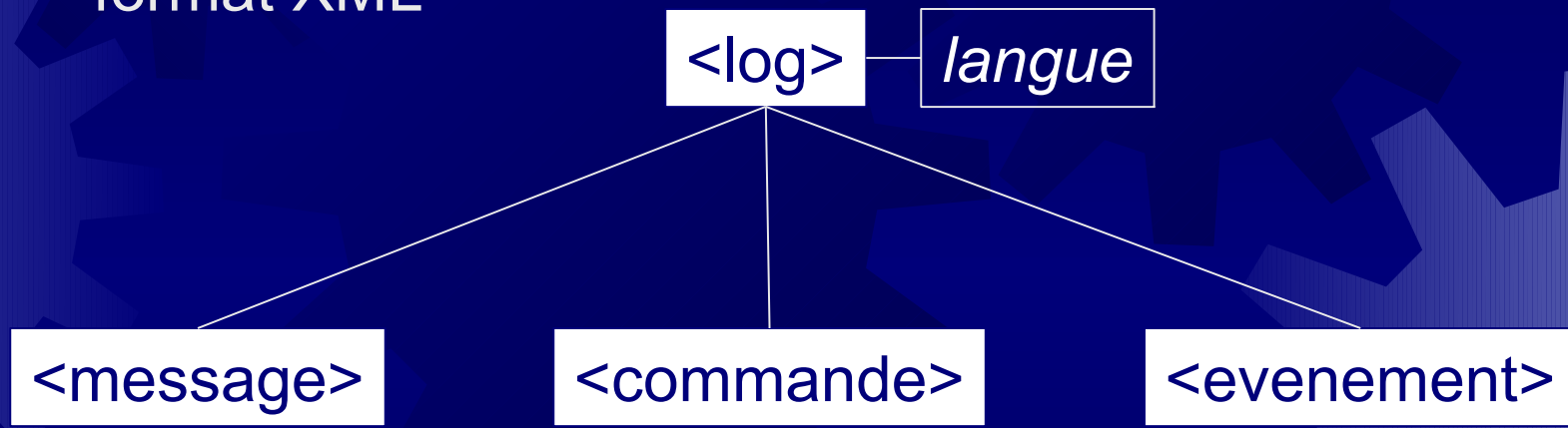
- ▶ Structure du corpus
 - ▶ format XML

<log>

langue

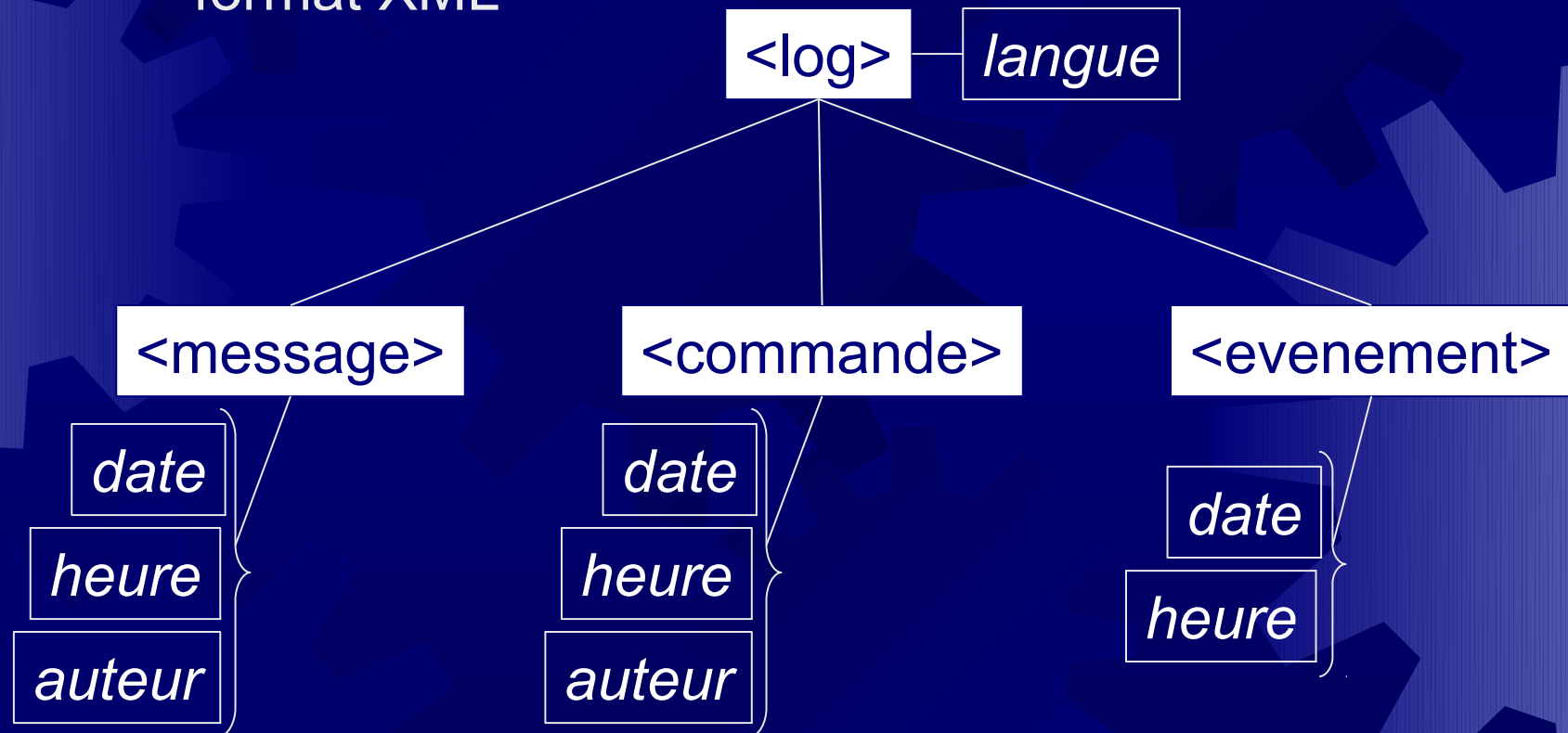
Un corpus de tchat

- ▶ Structure du corpus
 - ▶ format XML



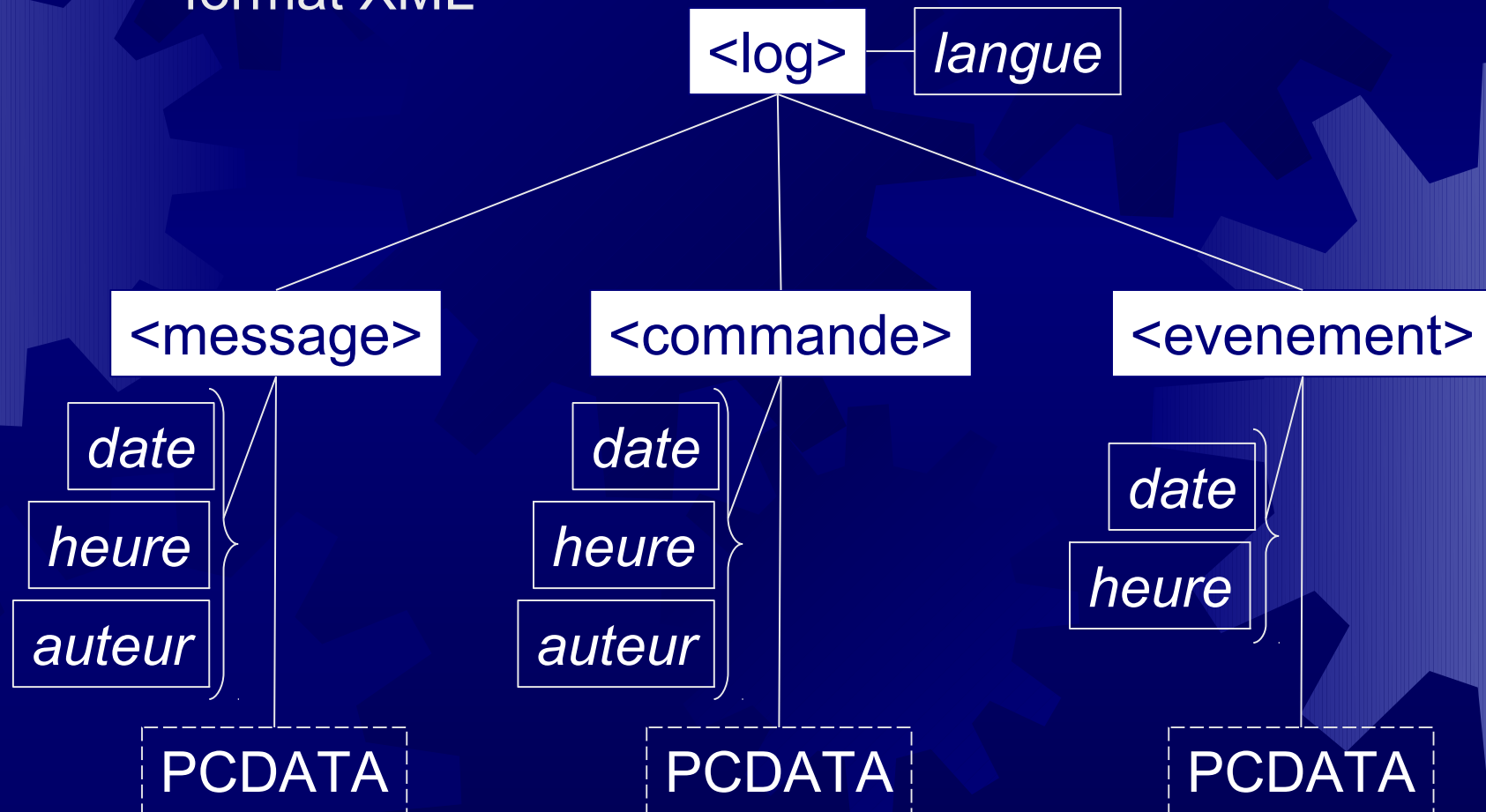
Un corpus de tchat

- Structure du corpus
 - format XML



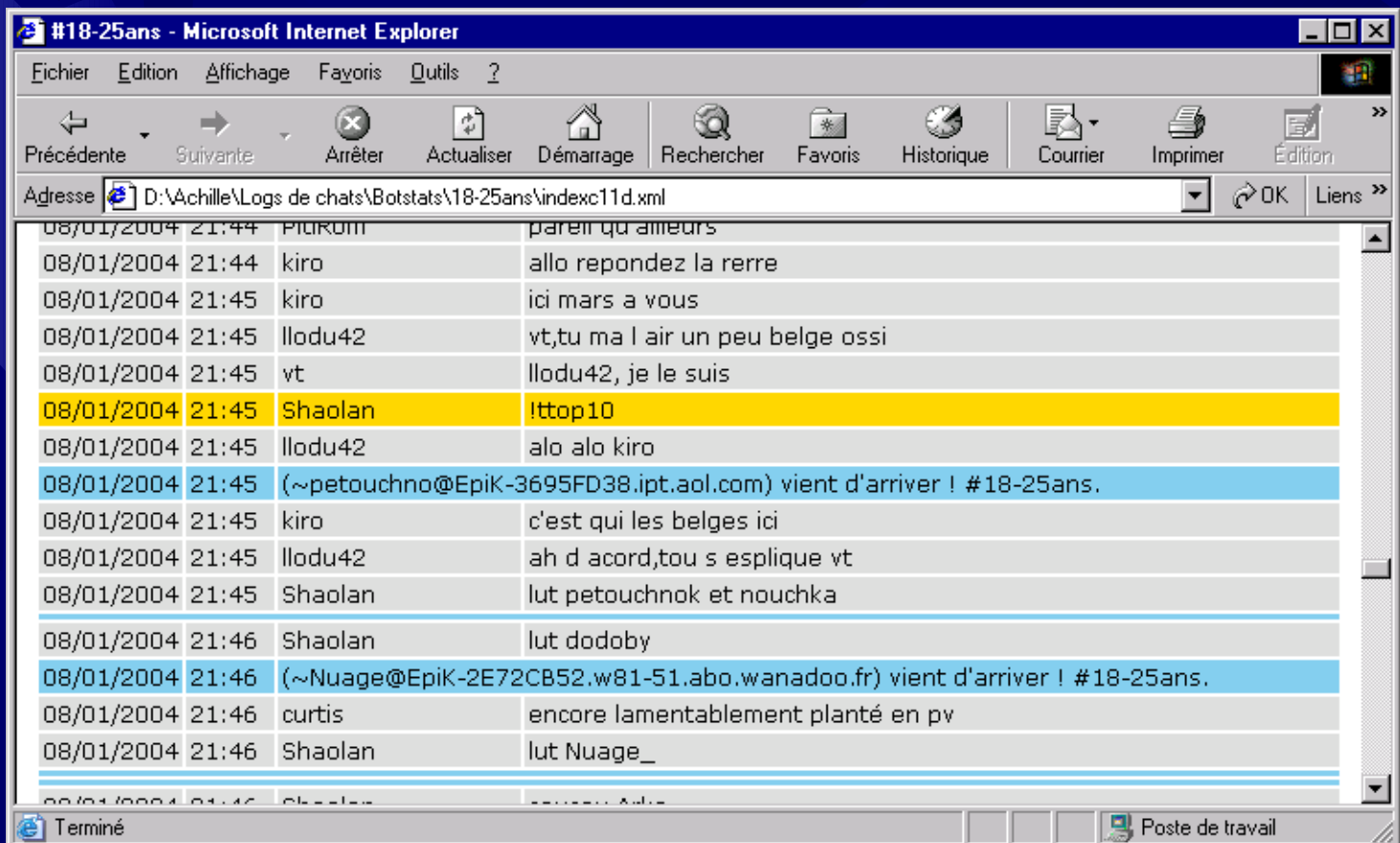
Un corpus de tchat

- Structure du corpus
 - format XML



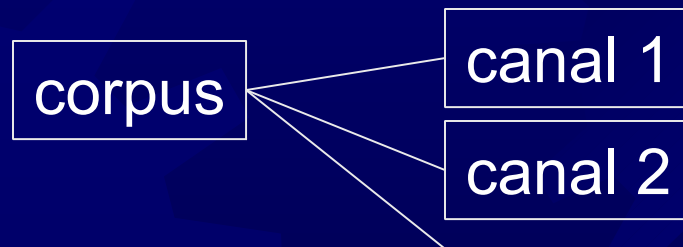
Un corpus de tchat

- ▶ Visualisation du corpus
- ▶ à l'aide d'une feuille de style XSLT



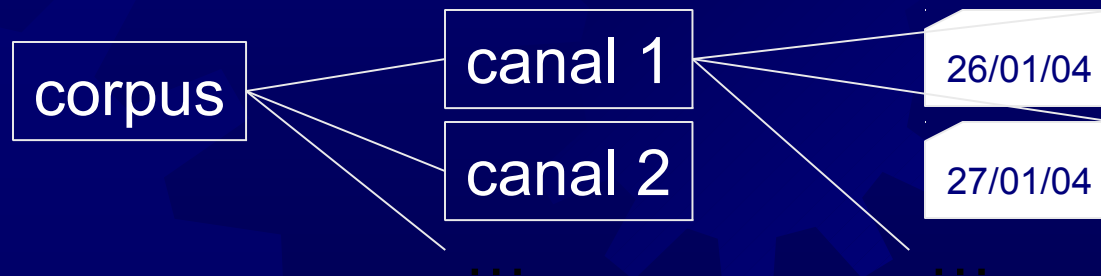
Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



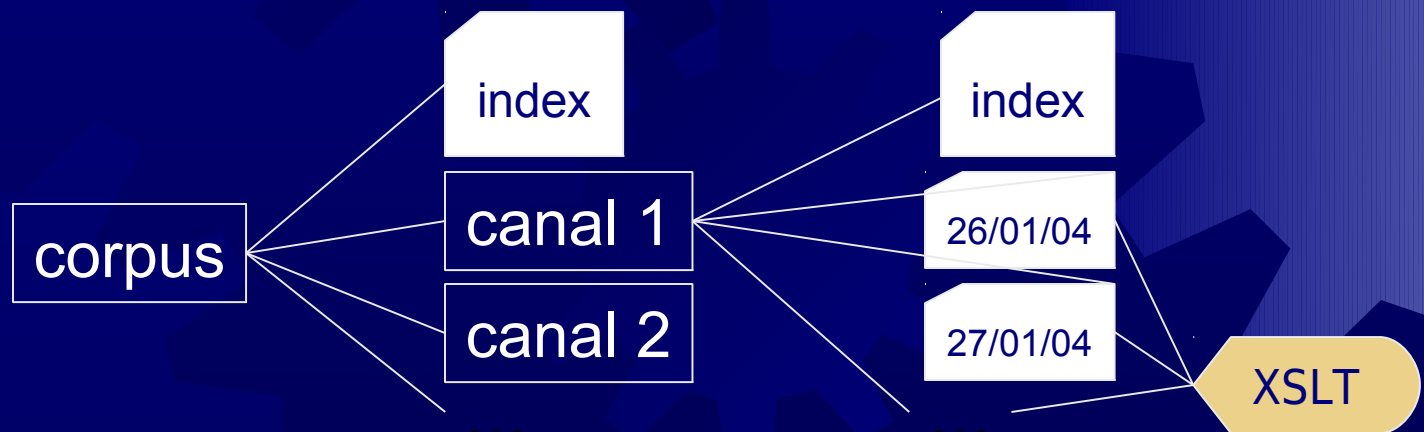
Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



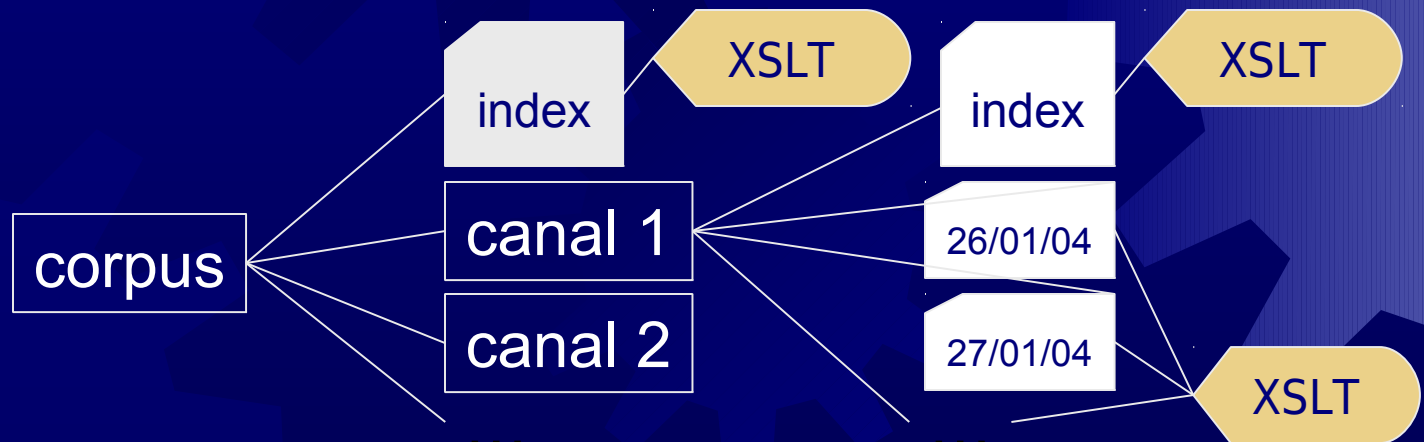
Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



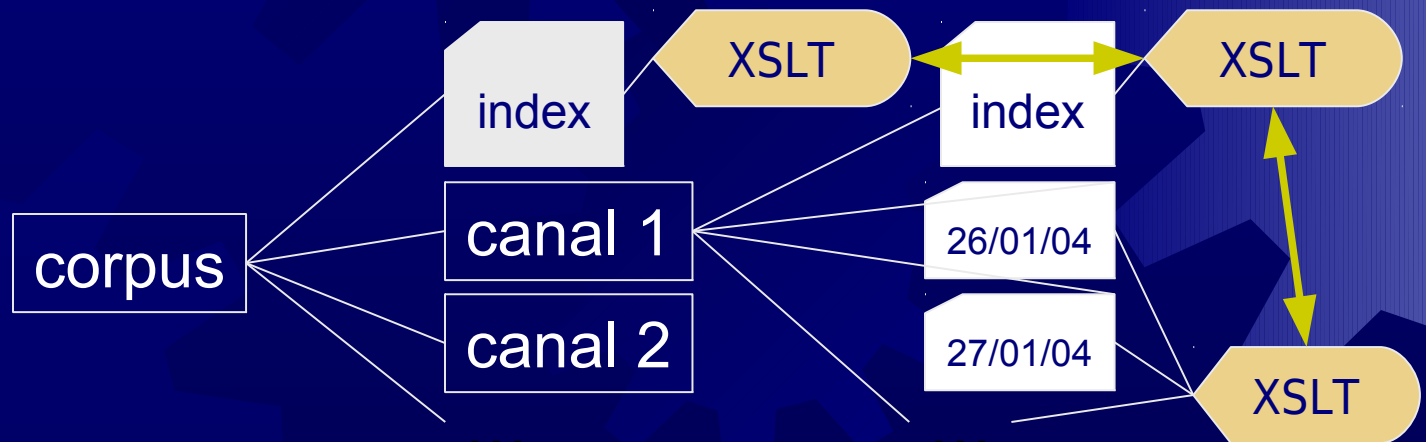
Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



Un corpus de tchat

- ▶ Organisation du corpus
 - ▶ à chaque canal correspond un répertoire
 - ▶ à chaque journée correspond un fichier
- ▶ Navigation dans le corpus
 - ▶ fichiers d'assemblage XML
 - ▶ visualisables grâce une feuille de style XSLT



La langue du tchat

- [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 :)

La langue du tchat

[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 :)

Abréviations

La langue du tchat

[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 :)

Fautes de
frappe

La langue du tchat

[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 :)

Fautes
d'orthographe

La langue du tchat

[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 :)

Fautes
d'orthographe

Corrections

La langue du tchat

[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 :)

Personnalisation

La langue du tchat

[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 [:)]

Emoticons

La langue du tchat

- [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 :)

Le tchat multilingue

- ▶ Qu'est-ce que c'est ?
 - ▶ un tchat dans lequel les utilisateurs conversent dans différentes langues
 - ▶ chaque utilisateur voit les messages traduits dans sa propre langue

Le tchat multilingue

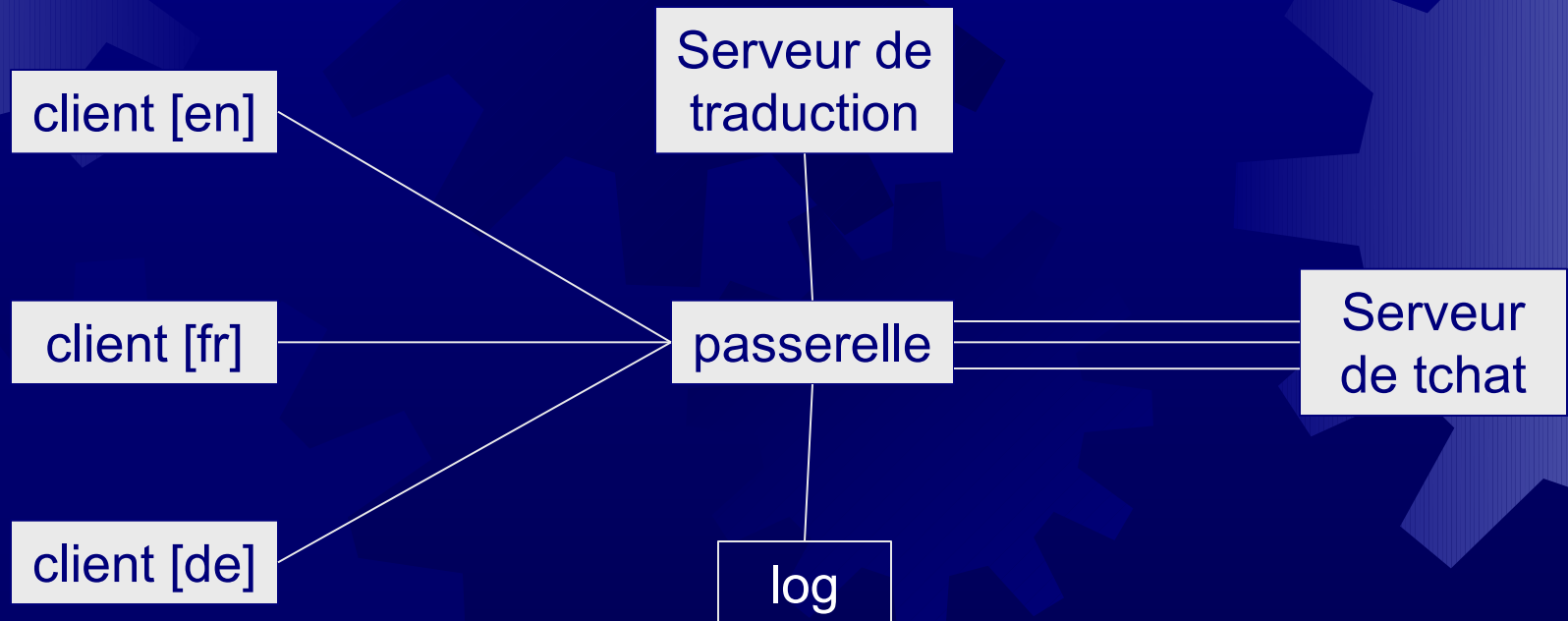
- ▶ Des applications existent
 - ▶ par exemple WorldLingo Chat
 - ▶ [captures d'écran]

Le tchat multilingue

- ▶ Mais ces applications sont encore très peu utilisées
 - ▶ pas de prise en compte des spécificités de la langue du tchat
 - ▶ TA « classique » de mauvaise qualité
- ▶ Comment améliorer ces applications ?
 - ▶ déterminer précisément quels problèmes pose la TA de tchat « classique »
 - ▶ pour cela, constituer un corpus de tchat multilingue par TA « classique »
- ▶ Ce que l'on voudrait évaluer
 - ▶ à quel point l'utilisateur acceptera-t-il de normaliser ?
 - ▶ à quel point une « mauvaise » TA est-elle utilisable ?

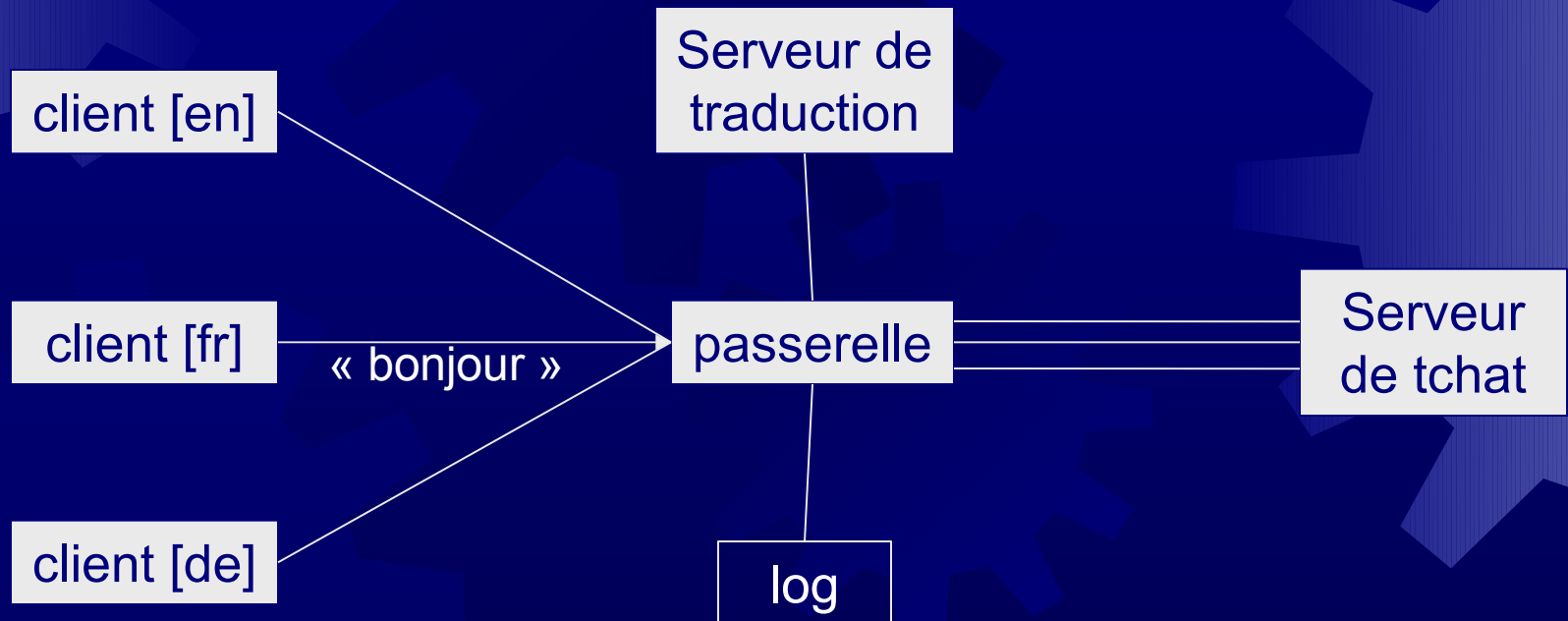
Système de collecte

- ▶ Pour construire notre corpus
 - ▶ on a besoin d'un système de tchat multilingue avec TA « classique »
- ▶ Système de type « passerelle »



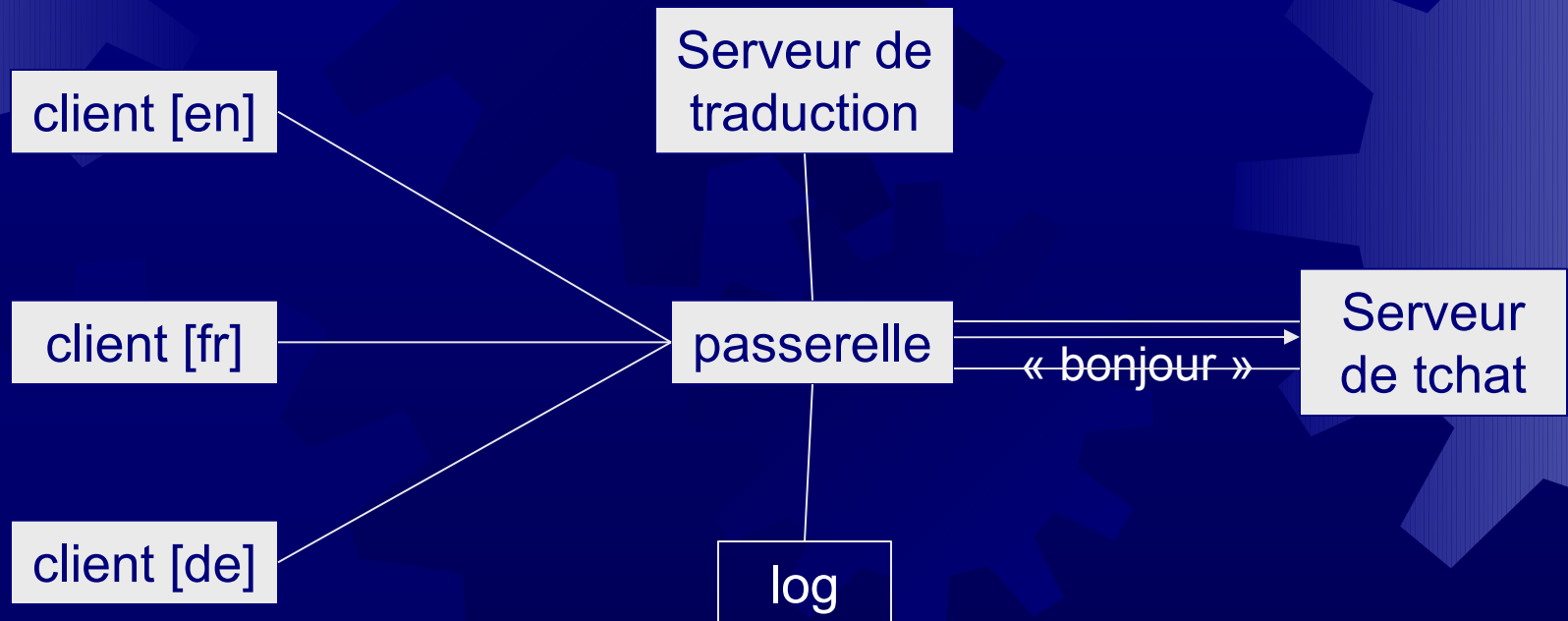
Système de collecte

- Pour construire notre corpus
 - on a besoin d'un système de tchat multilingue avec TA « classique »
- Système de type « passerelle »



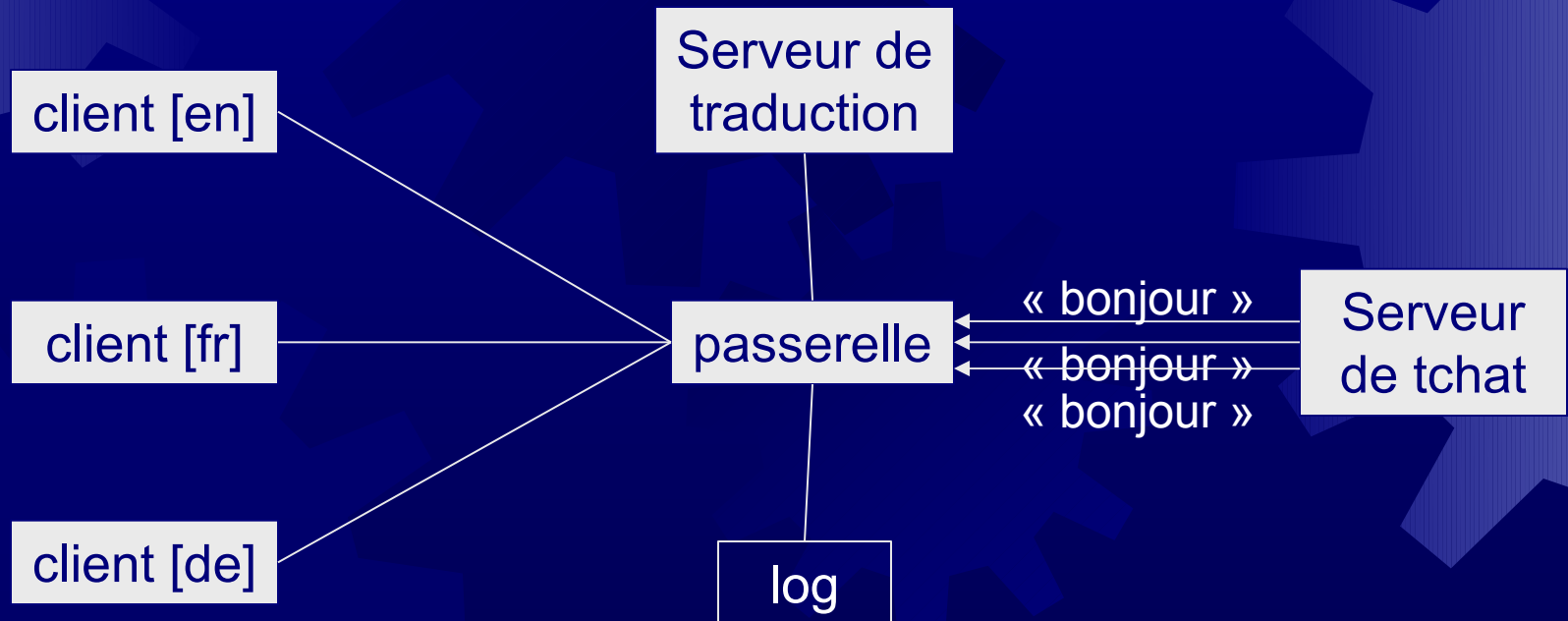
Système de collecte

- ▶ Pour construire notre corpus
 - ▶ on a besoin d'un système de tchat multilingue avec TA « classique »
- ▶ Système de type « passerelle »



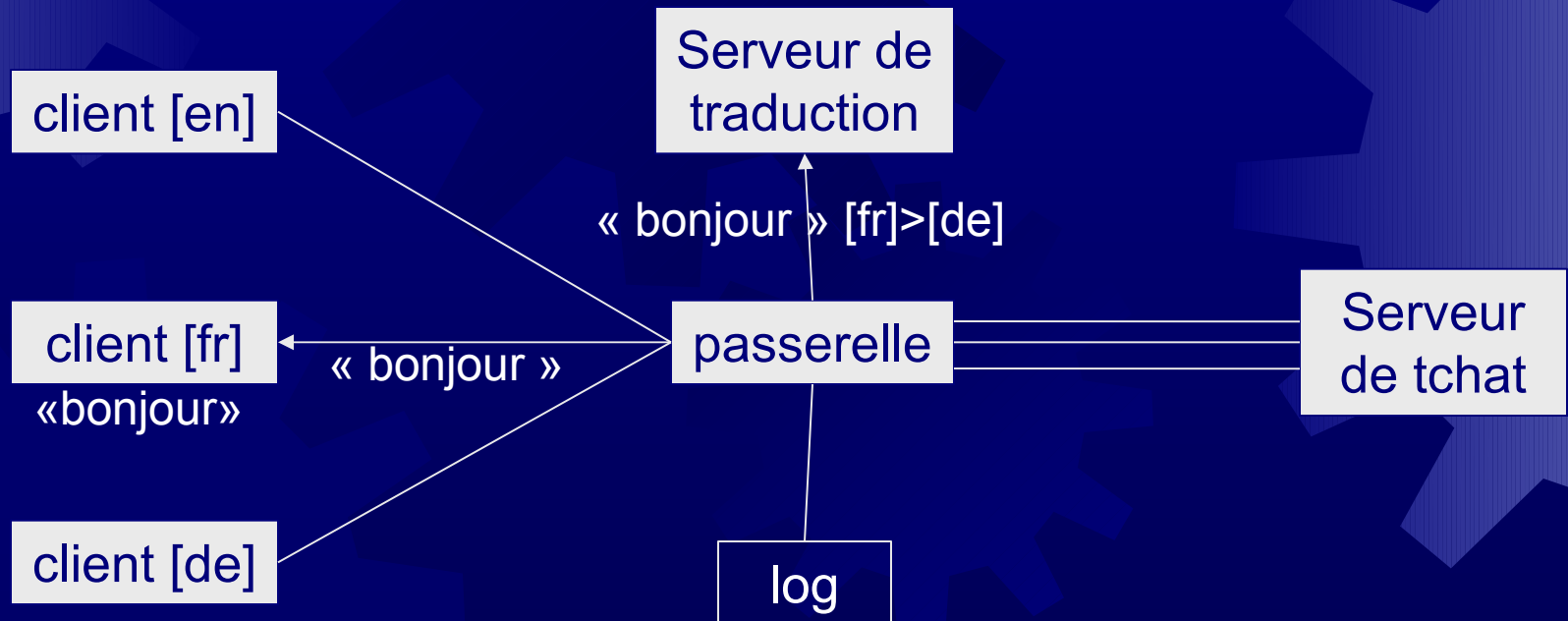
Système de collecte

- Pour construire notre corpus
 - on a besoin d'un système de tchat multilingue avec TA « classique »
- Système de type « passerelle »



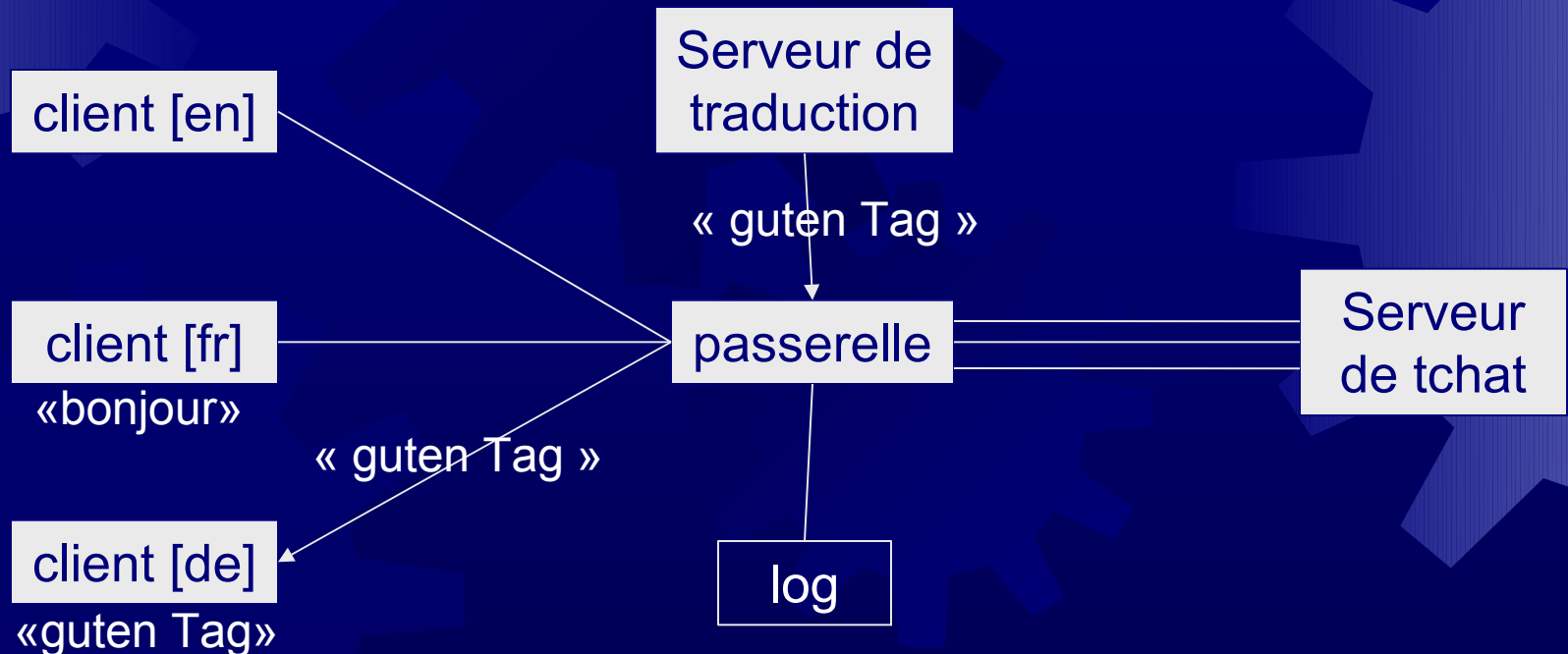
Système de collecte

- Pour construire notre corpus
 - on a besoin d'un système de tchat multilingue avec TA « classique »
- Système de type « passerelle »



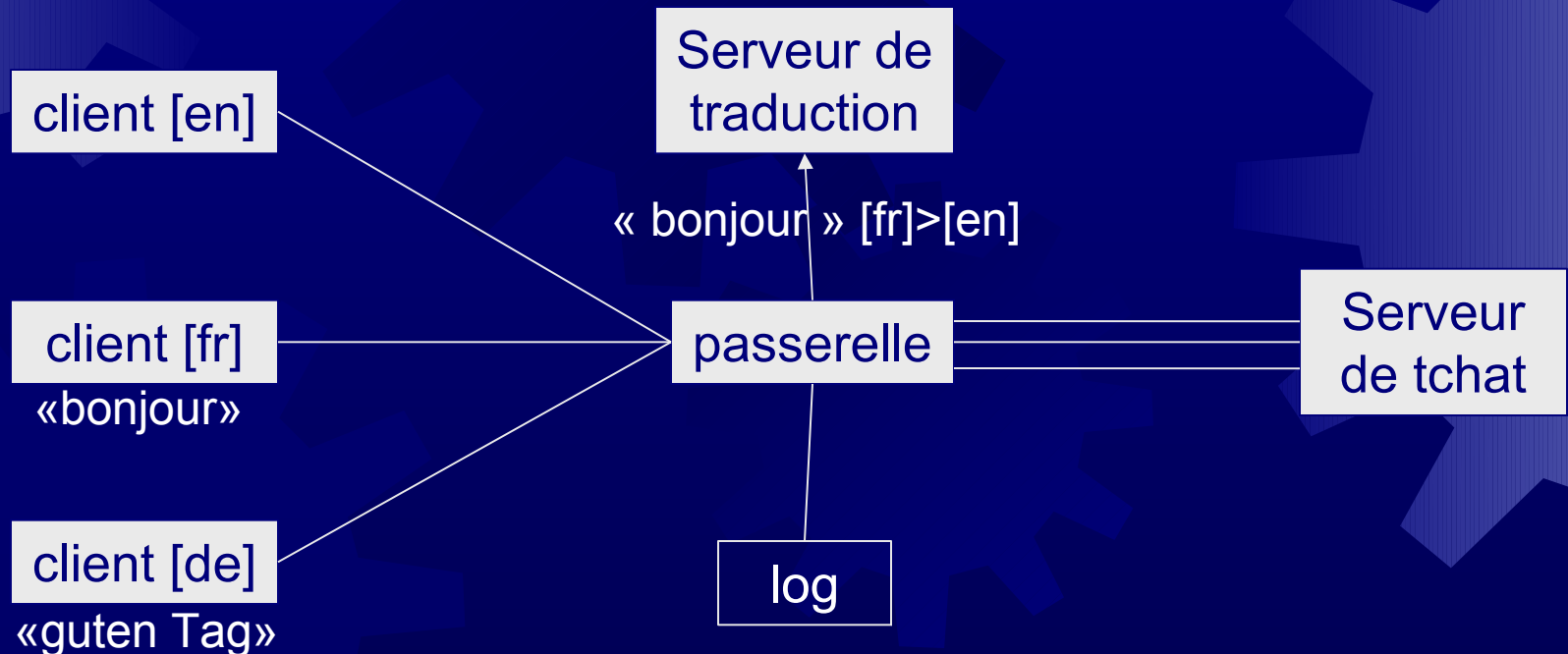
Systeme de collecte

- Pour construire notre corpus
 - on a besoin d'un système de tchat multilingue avec TA « classique »
- Systeme de type « passerelle »



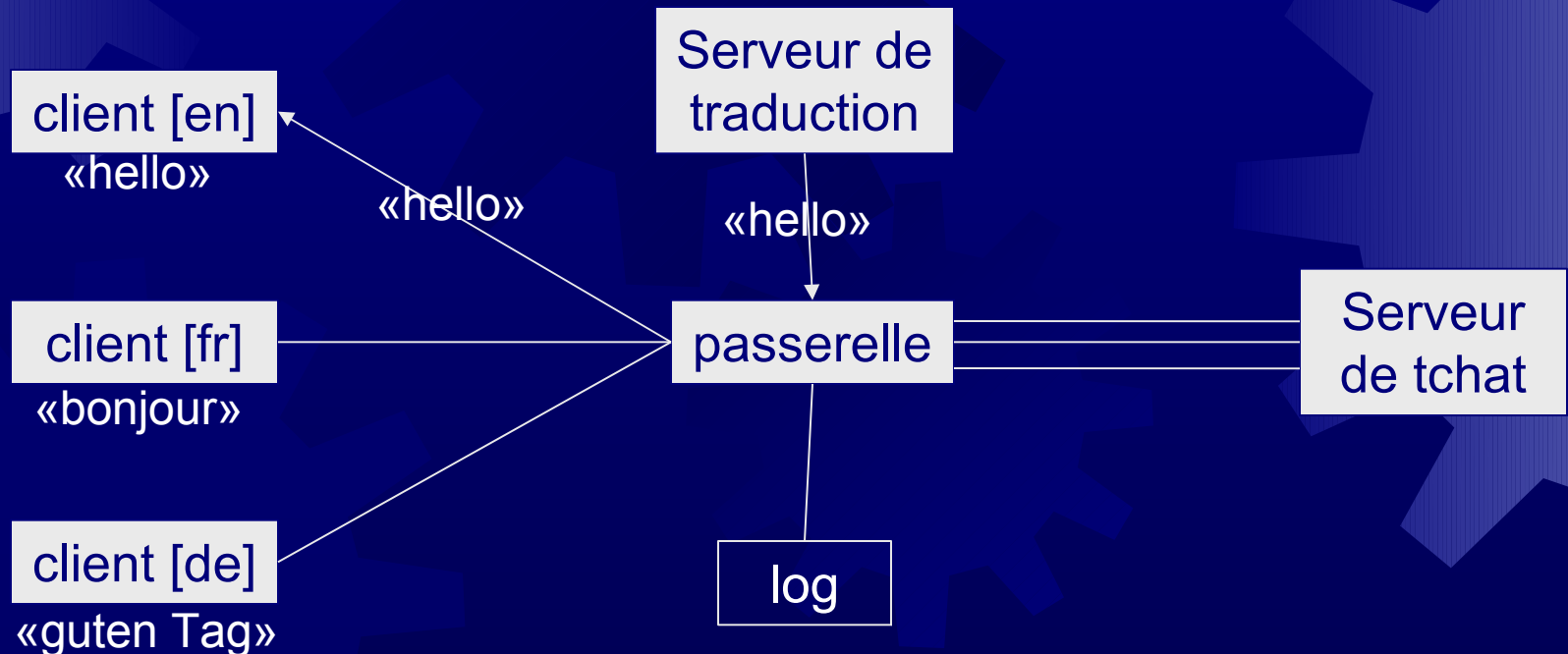
Système de collecte

- ▶ Pour construire notre corpus
 - ▶ on a besoin d'un système de tchat multilingue avec TA « classique »
- ▶ Système de type « passerelle »



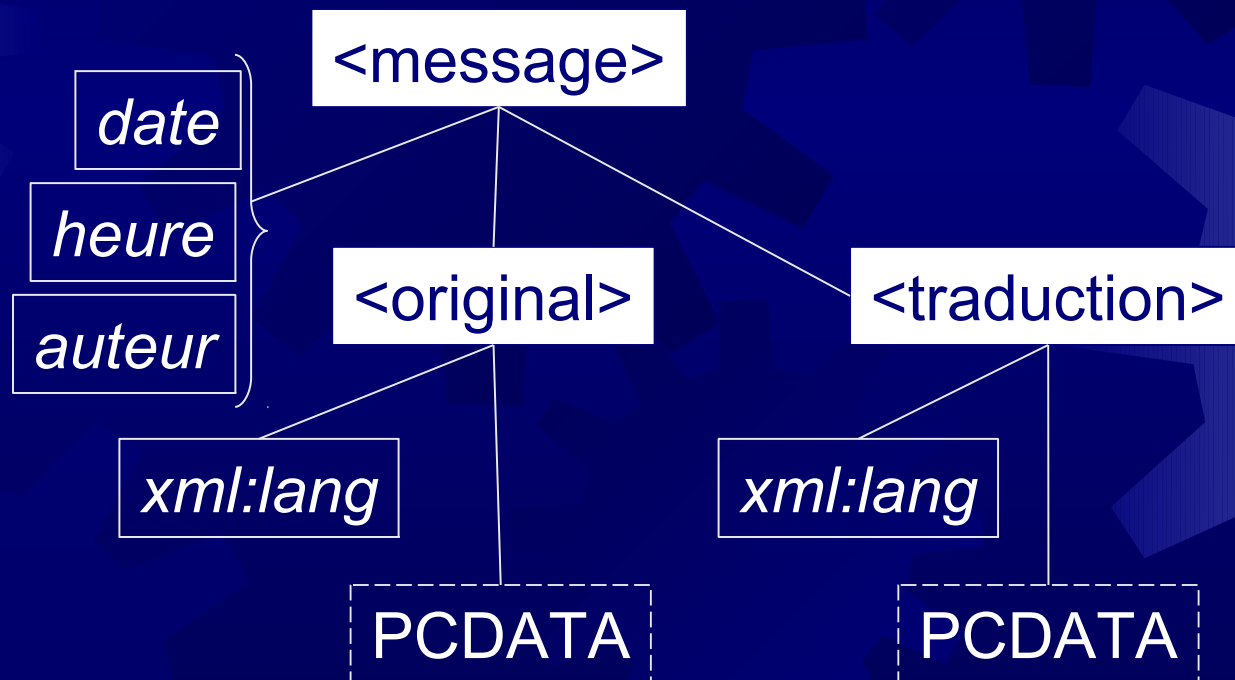
Systeme de collecte

- Pour construire notre corpus
 - on a besoin d'un système de tchat multilingue avec TA « classique »
- Systeme de type « passerelle »



Un corpus multilingue

- ▶ Structure du corpus
 - ▶ format XML très proche de celui du corpus monolingue
 - ▶ polyphrases pour les messages



- ▶ informations linguistiques sur les utilisateurs

Conclusion

► Résultats

- un corpus de tchat monolingue
- le tchat : une modalité spécifique, entre langue écrite et langue parlée
- un système de collecte de tchat multilingue

► Perspectives

- thèse CIFRE avec Prosodie, portant sur la vidéoconférence multilingue assistée par ordinateur