

RECITAL 2005

Constitution d'un corpus de français tchaté



GETA
CLIPS

Achille Falaise
achille.falaise@imag.fr
Doctorant CIFRE





Introduction

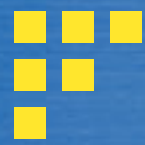
■ Contexte

- Intérêt grandissant pour le tchat
 - Proche de la langue parlée
 - Dialogue, interactions
 - Naissance et évolution de communautés linguistiques
 - Tchat et messagerie instantanée : 150 millions d'utilisateurs en 2001, et commence à gagner le monde de l'entreprise
- Manque de ressources facilement accessibles
 - Eulogos [it] (850 000 mots)
 - Sites de logs
- Printemps 2004: stage de M2R sur le tchat
 - Collecte et formatage d'un corpus



Introduction

- Intérêt du corpus collecté
 - Volumineux (23 millions de mots)
 - Thématiques variées (105 canaux)
 - Légalement distribuable
- Publication du corpus
 - Afin de réduire le travail de collecte pour les autres travaux en rapport le tchat
 - Consultable via une interface web



Plan

- Qu'est-ce que le tchat ?
- La langue du tchat
 - Lexique
 - Syntaxe
- Le corpus
 - Présentation
 - Structure
 - Publication
- Conclusion & perspectives



Qu'est-ce que le tchat ?

- Un outil de dialogue écrit en ligne
 - En quasi temps-réel (≠ courriel, forum, newsgroup)
 - Gratuit (≠ SMS)
 - Public (≠ messagerie instantanée)
- Conversations au sein de canaux
- Usages
 - Récréatif
 - Technique
 - Professionnel
- Robots
 - Administration, statistiques...
 - Quizz, pendu...
 - Dépêches AFP



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Qu'est-ce que le tchat ?

Exemple

A> soir tlm
A> kikooooooooo B :*)
B> kikooo A :)
<C vient de se connecter>
A> yo C
C> yop all
C> yop A
A> o y repond today
C> kikouuu B
C> yop midam B
C> yop D
A> j'ai de la chance loll



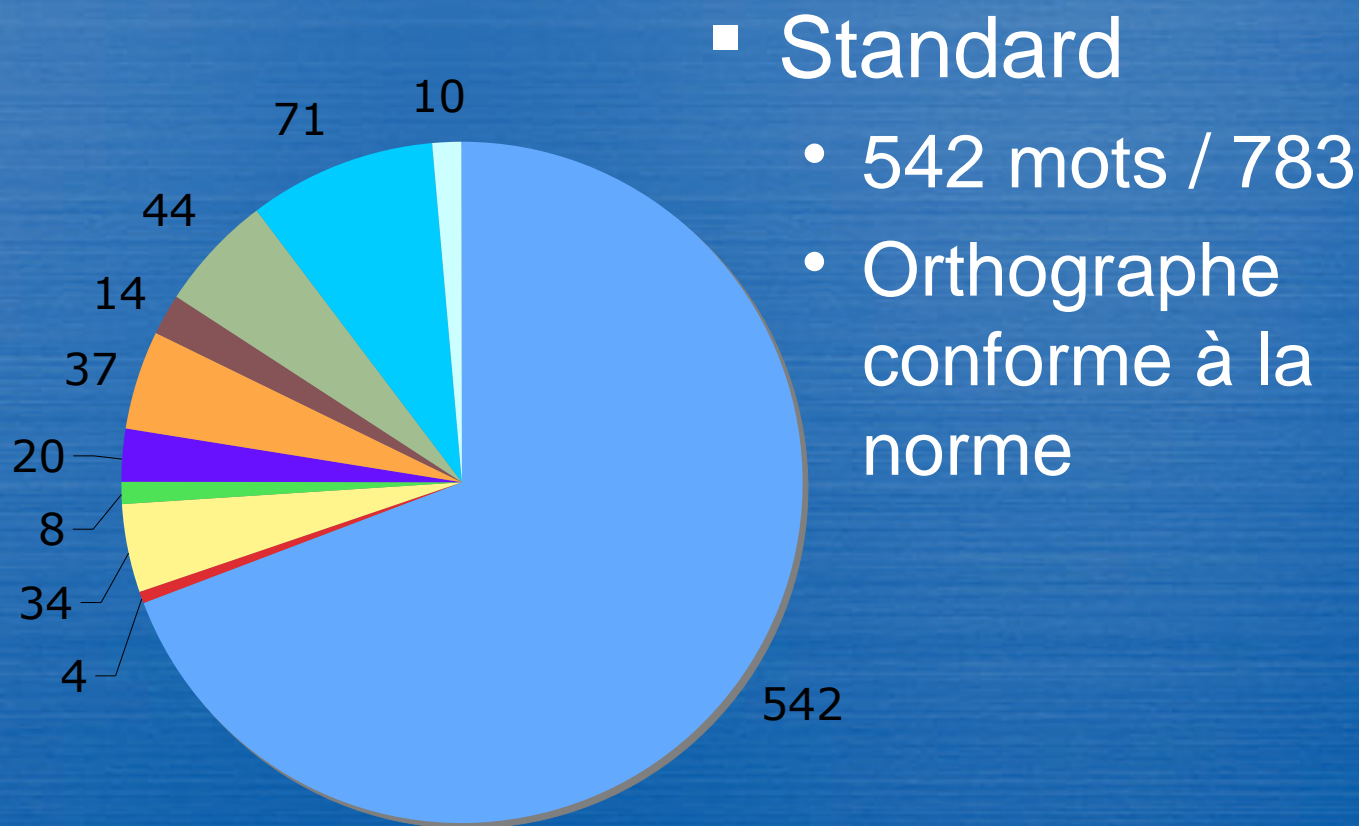
La langue du tchat

Phénomènes lexicaux

- Présentation de phénomènes lexicaux caractéristiques du tchat
- Quantification grossière de ces phénomènes à partir de 200 énoncés (783 mots) extraits aléatoirement du canal #18-25ans

Aperçu du français tchaté

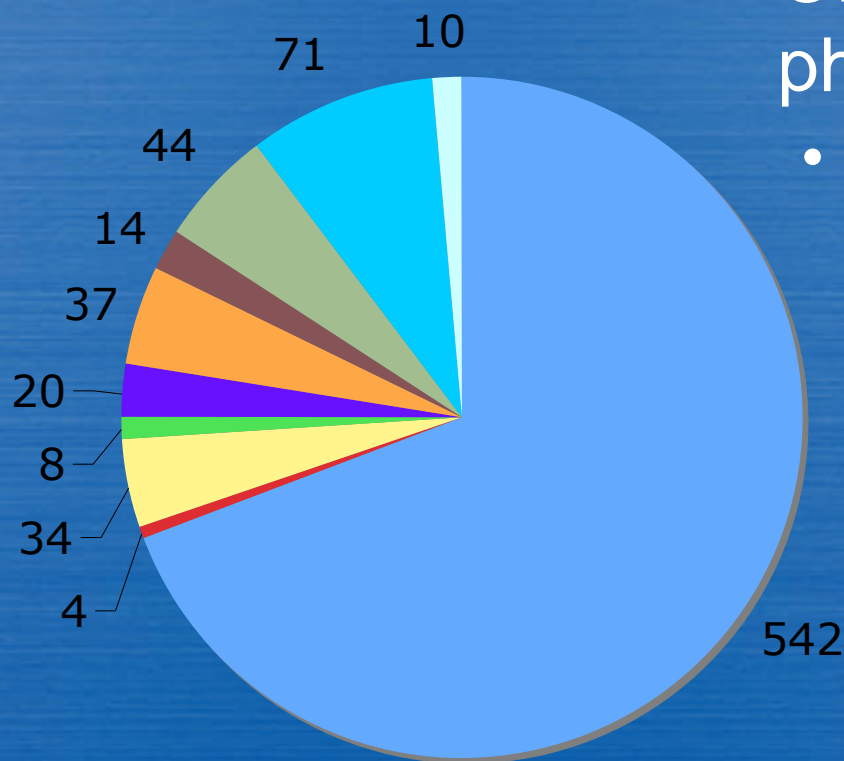
Phénomènes lexicaux



783 mots sélectionnés aléatoirement sur le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux



■ Orthographe phonétique

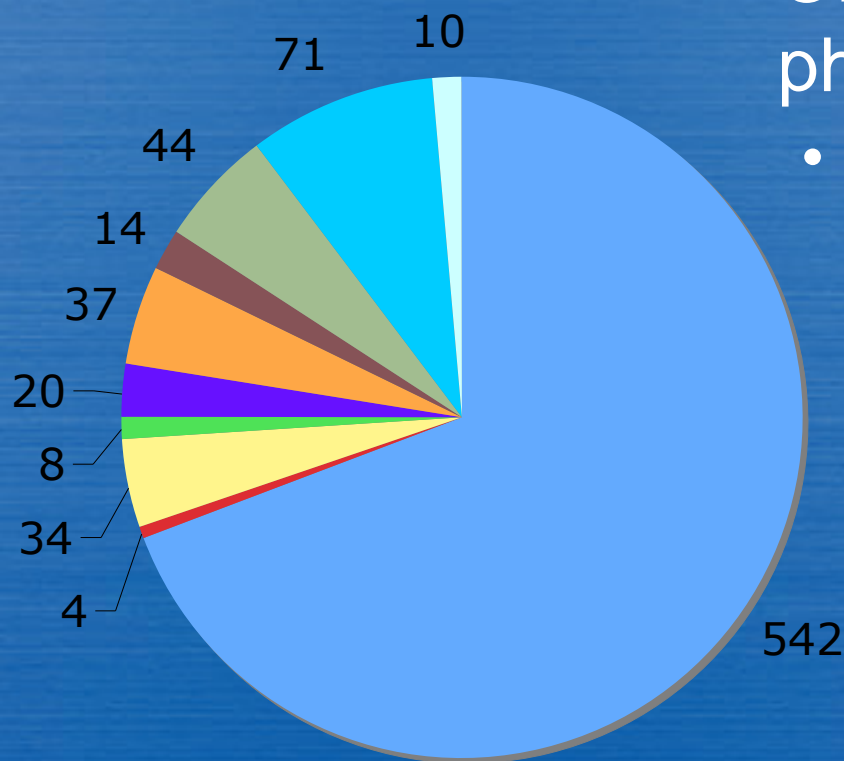
• 71 mots / 783

- *c ce ke g eu*
- *kikoo*
- *kikooooo*
- *oki*
- *po moi*
- *2m1*
- *dem1*
- *2main*

Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux



Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Orthographe phonétique

- 71 mots / 783

- *c ce ke g eu*

- *kikoo*

- *kikooooo*

- *oki*

- *po moi*

- *2m1*

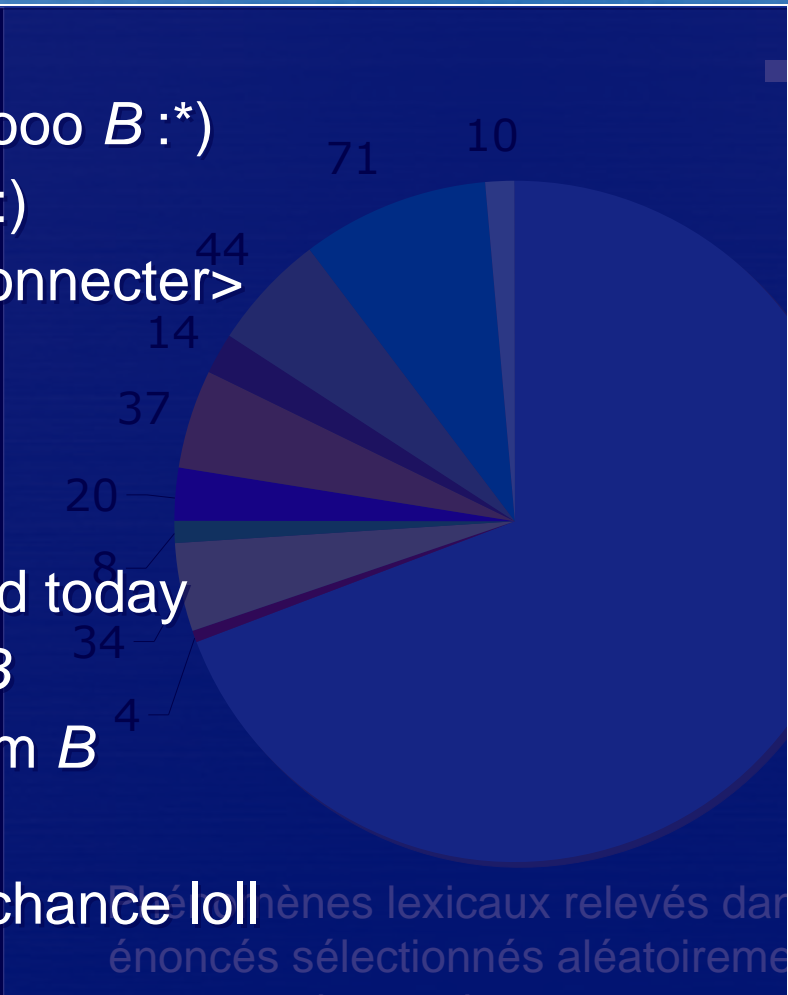
- *dem1*

- *2main*

Aperçu du français tchaté

Phénomènes lexicaux

A> soir tlm
 A> kikooooooooo B :*)
 B> kikooo A :)
 <C vient de se connecter>
 A> yo C
 C> yop all
 C> yop A
 A> o y repond today
 C> **kikouuu** B
 C> yop midam B
 C> yop D
 A> j'ai de la chance lolll



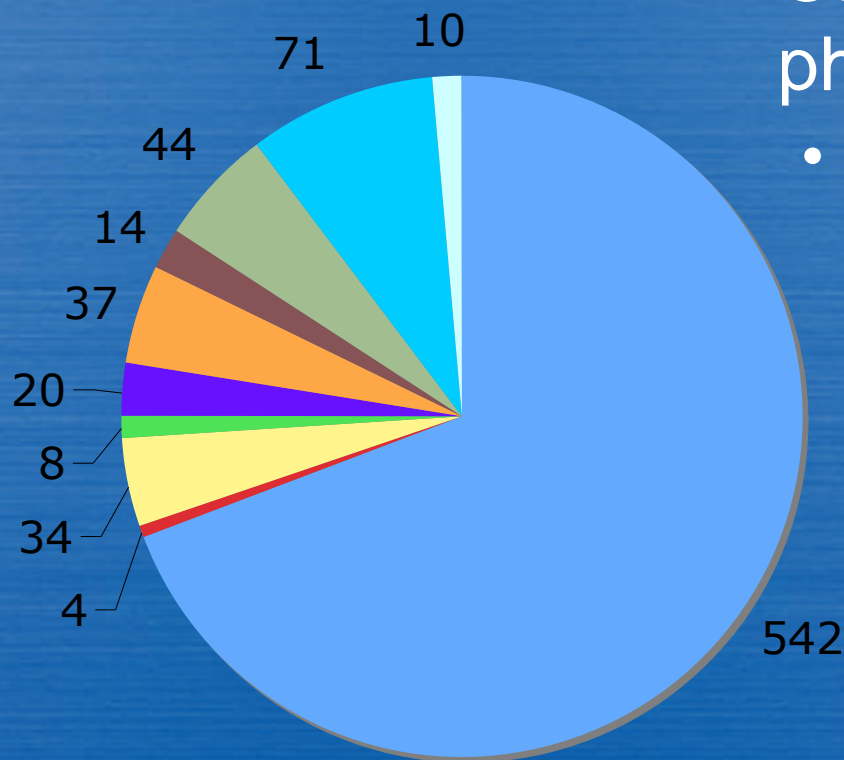
Orthographe phonétique

- 71 mots / 783
 - c ce ke g eu
 - kikoo
 - kikooooo
 - oki
 - po moi
 - 2m1
 - dem1
 - 2main

Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur

Aperçu du français tchaté

Phénomènes lexicaux



Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Orthographe phonétique

- 71 mots / 783

- *c ce ke g eu*

- *kikoo*

- *kikooooo*

- *oki*

- *po moi*

- *2m1*

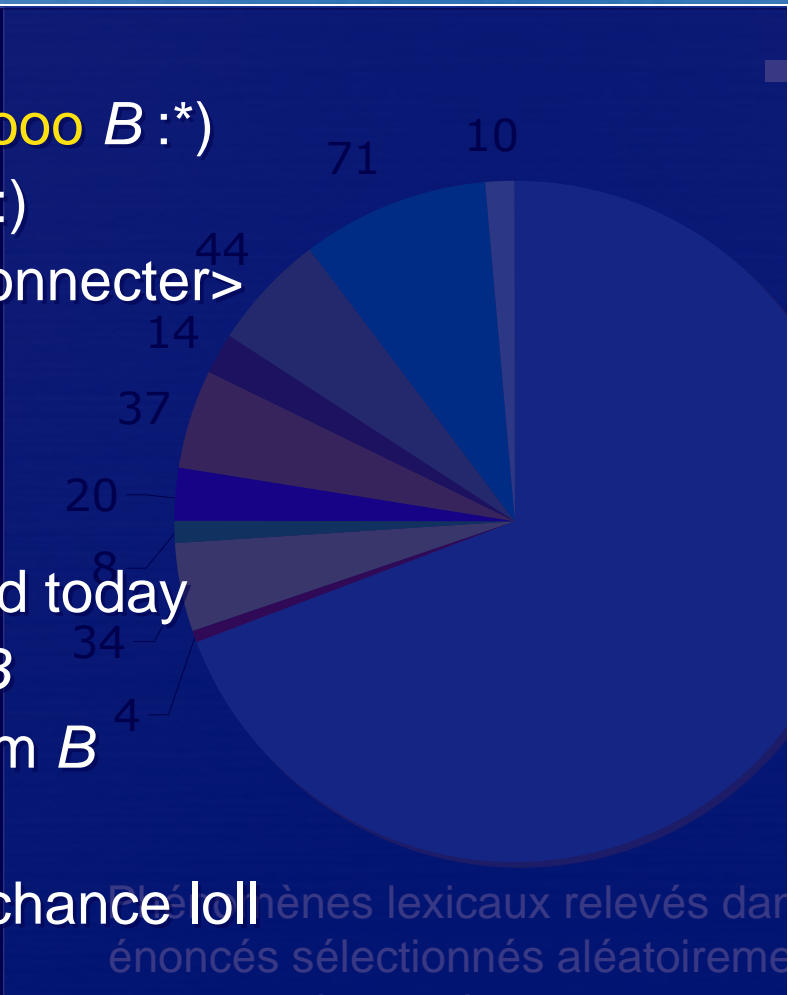
- *dem1*

- *2main*

Aperçu du français tchaté

Phénomènes lexicaux

A> soir tlm
 A> kikooooooooo B :*)
 B> kikooo A :)
 <C vient de se connecter>
 A> yo C
 C> yop all
 C> yop A
 A> o y repond today
 C> kikouuu B
 C> yop midam B
 C> yop D
 A> j'ai de la chance lolll



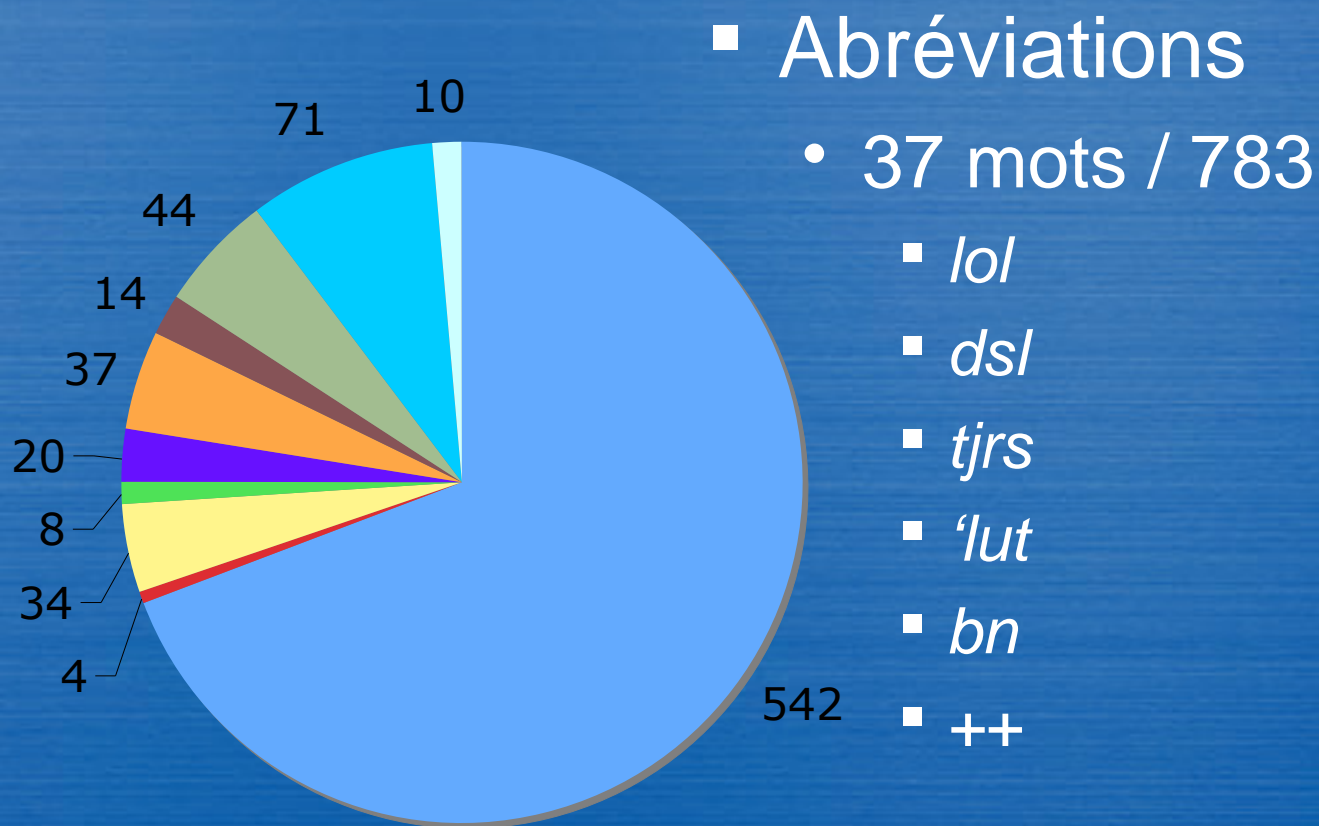
Orthographe phonétique

- 71 mots / 783
 - c ce ke g eu
 - kikoo
 - kikooooo
 - oki
 - po moi
 - 2m1
 - dem1
 - 2main

Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux



Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux

A> soir **tlm**

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

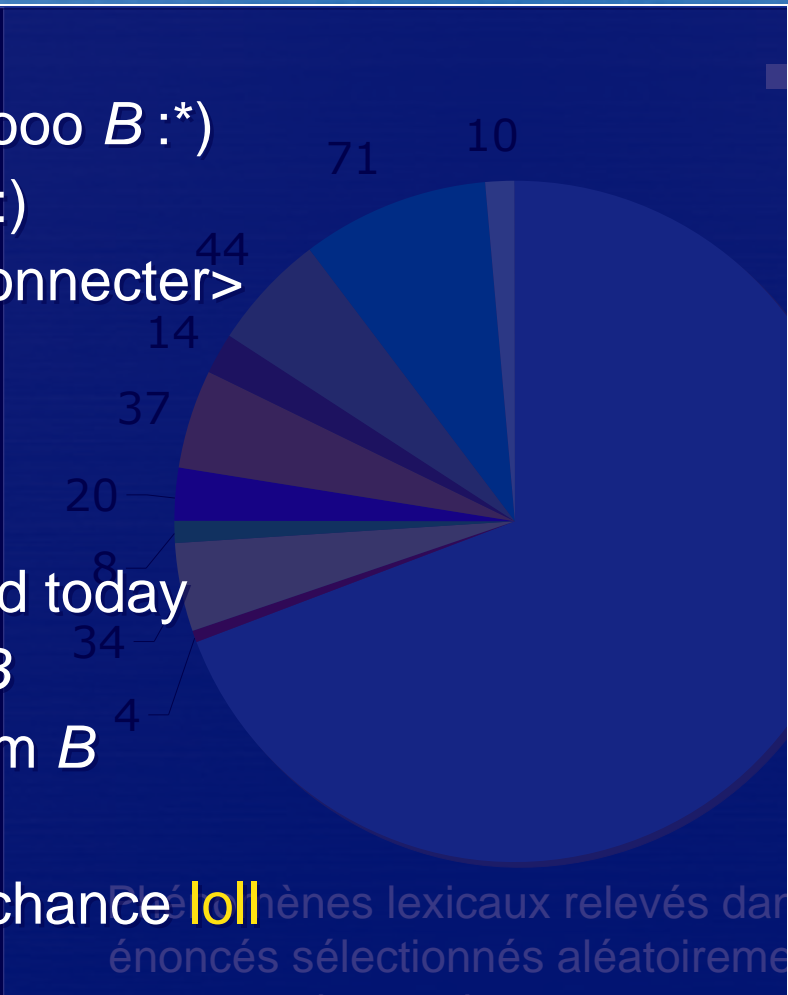
A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance **loll**



Abréviations

- 37 mots / 783
- lol
- dsl
- tjrs
- 'lut
- bn
- ++

Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur

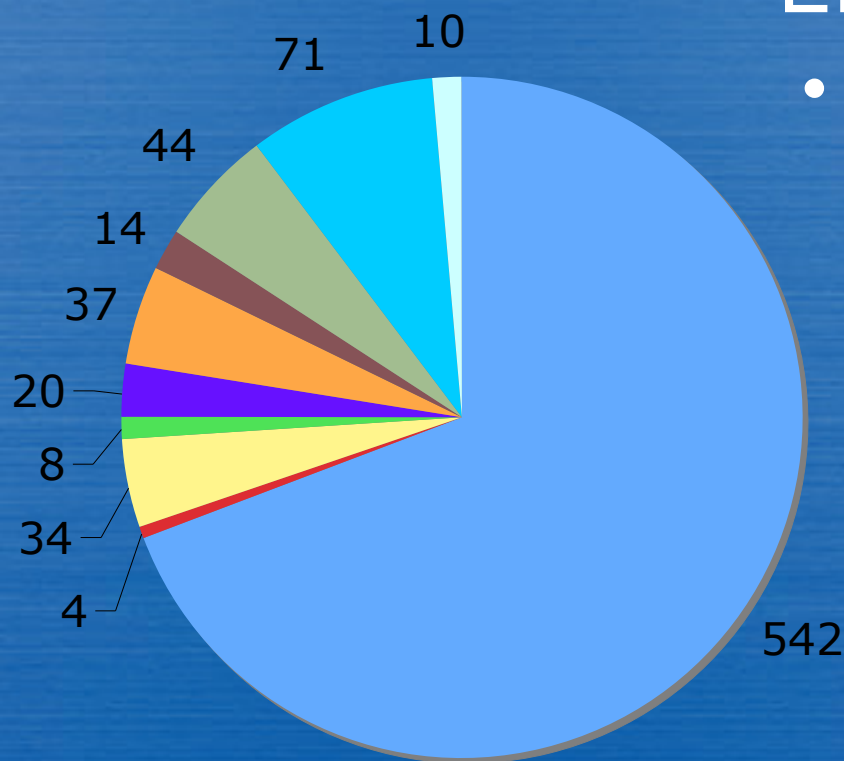
Aperçu du français tchaté

Phénomènes lexicaux

■ Emoticons

• 34 mots / 783

- ;)
- :)
- =))
- :-*
- :p



Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux

A> soir tlm

A> kikooooooooo B :*)

B> kikooo A :)

<C vient de se connecter>

A> yo C

C> yop all

C> yop A

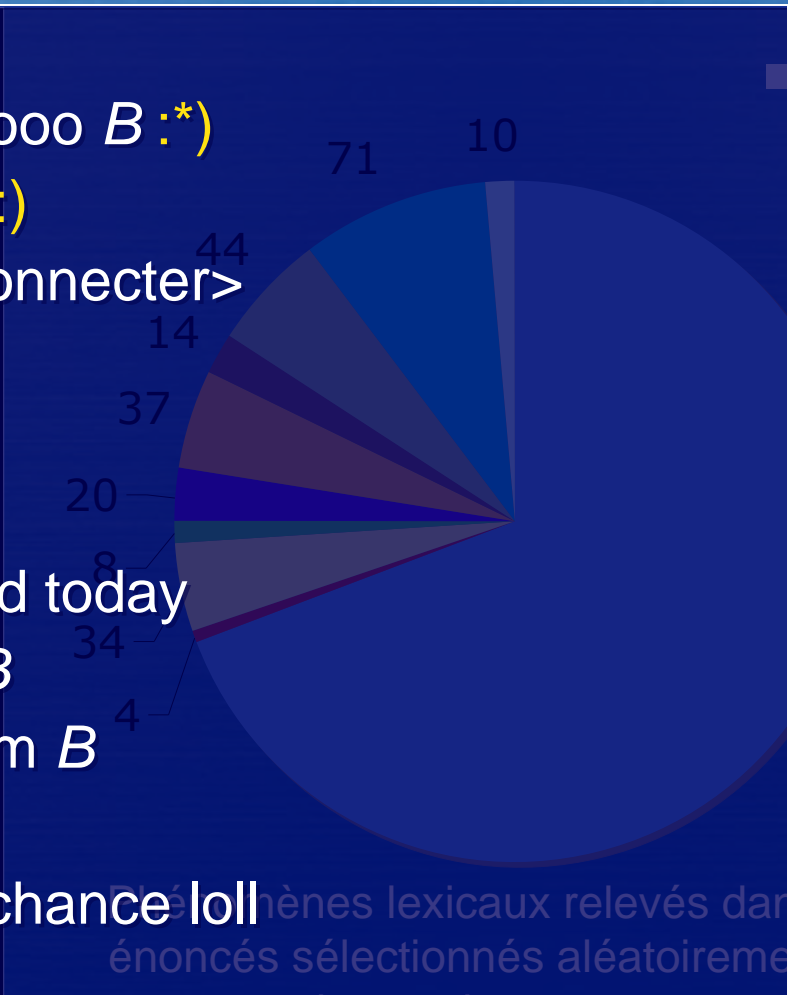
A> o y repond today

C> kikouuu B

C> yop midam B

C> yop D

A> j'ai de la chance lolll



Emoticons

• 34 mots / 783

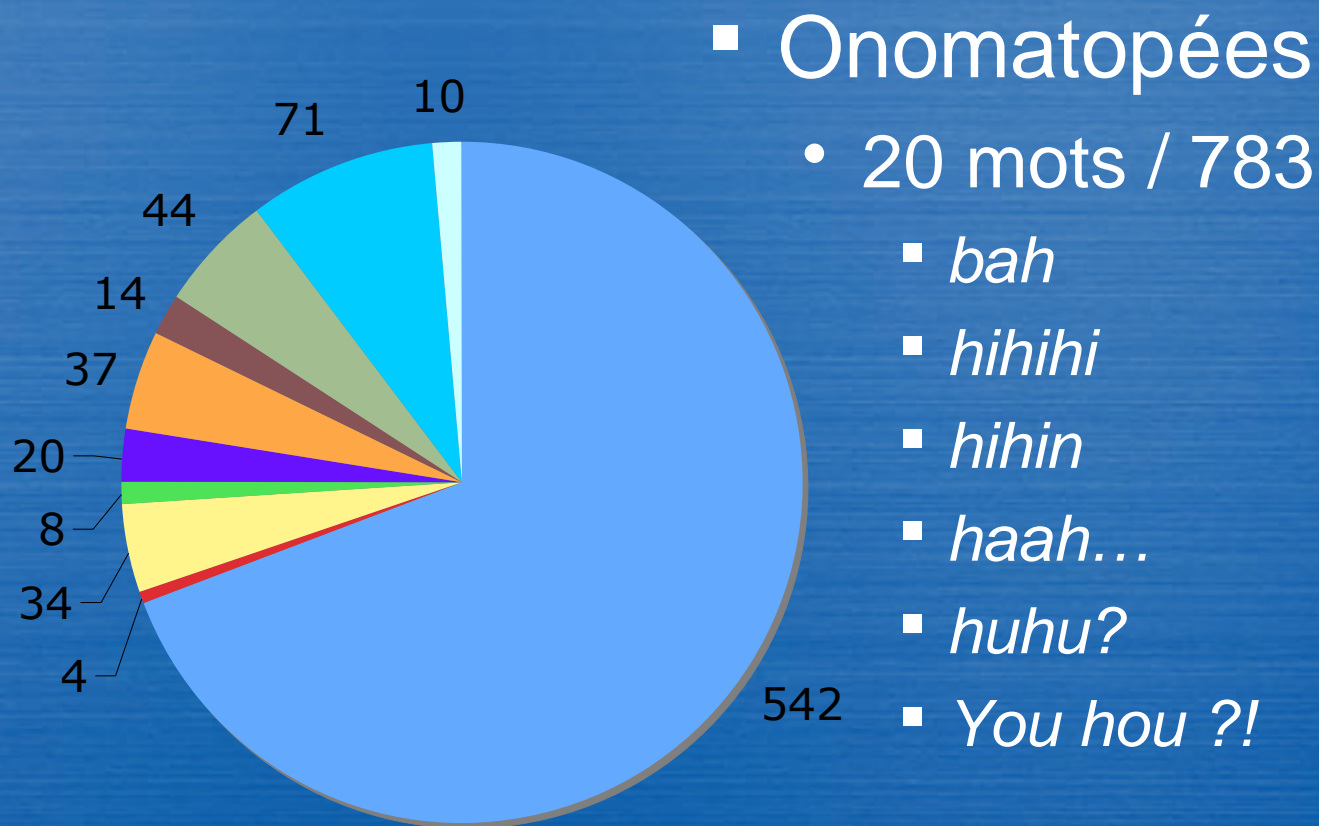
- ;)
- :)
- =))
- :-*
- :p

Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur

le canal #18-25ans

Aperçu du français tchaté

Phénomènes lexicaux

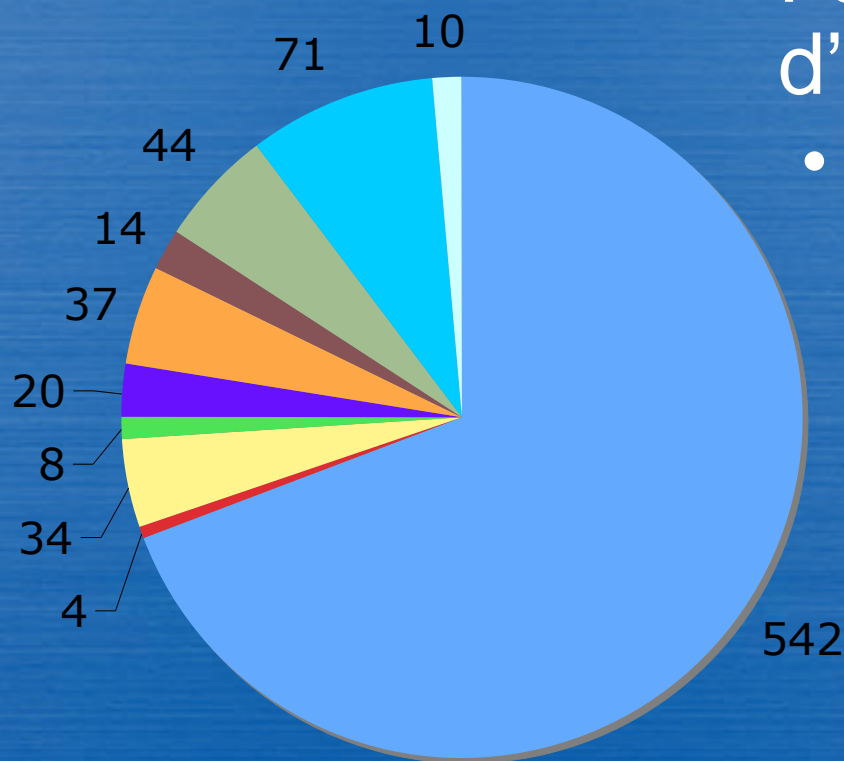


Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

- Onomatopées
 - 20 mots / 783
 - *bah*
 - *hihihi*
 - *hihin*
 - *haah...*
 - *huhu?*
 - *You hou ?!*

Aperçu du français tchaté

Phénomènes lexicaux



Phénomènes lexicaux relevés dans 200 énoncés sélectionnés aléatoirement sur le canal #18-25ans

- Fautes d'orthographe
 - 10 mots / 783
 - je t'*envoi* rien en *plaine* tronche
 - facile a dire une fois *balancer*
 - j'y ai *jamasi* habité



Aperçu du français tchaté

Phénomènes syntaxiques

- A rapprocher de l'oral
 - Topicalisations
 - Découpage en groupes prosodiques
 - Lorsqu'un message est long, l'utilisateur le tronçonne souvent en plusieurs morceaux
 - La coupure se fait généralement sur les frontières de groupes prosodiques
- Corrections
 - Objectif de l'utilisateur: se faire comprendre
 - Prêt à faire un peu de post-édition si nécessaire

Aperçu du français tchaté

Phénomènes syntaxiques

■ A rapprocher de l'oral

- Topicalisations
- Découpage en groupes prosodiques
 - Lorsqu'un message est long, l'utilisateur le tronçonne souvent en plusieurs morceaux
 - La coupure se fait généralement sur les frontières de groupes prosodiques

■ Corrections

- Objectif de l'utilisateur: se faire comprendre
- Prêt à faire un peu de post-édition si nécessaire

A> B préfère sauce **tarat**

A> **tartare***

B> du tou

A> sisi

A> a point

C> poivre rulez

D> sauce bernaïse c la meilleure :)

A> frite moelleuse

C> nan a point sauce poivre

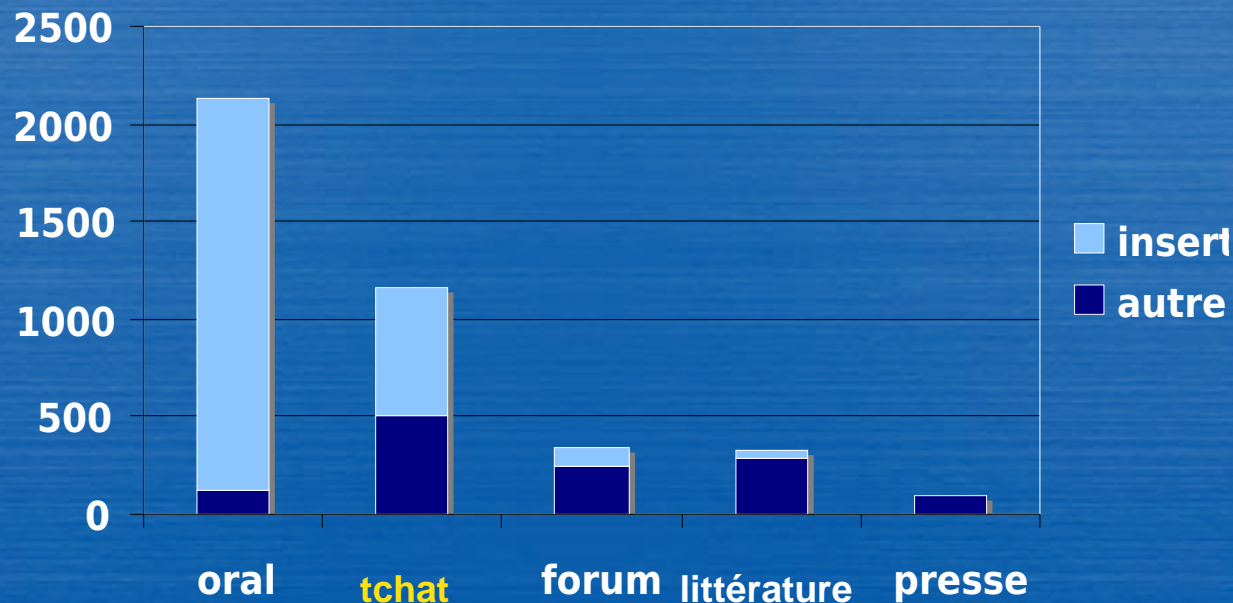
C> ça rox

A> lol

C> moelleuse mais croustillante stp

Aperçu du français tchaté

- Un autre indice de la validité d'un rapprochement avec l'oral
 - « L'analyse syntaxique de l'oral: problèmes et méthodes », C. Benzitoun et al. (TALN 2004): usages de « *bon* » comme insert dans 4 corpus de 440 000 mots
 - Comparaison avec un extrait de 35 000 mots issu du canal #18-25ans, et rapporté à la même échelle





Aperçu du français tchaté

- Caractéristiques de l'écrit tchaté:
 - Phénomènes lexicaux
 - Créativité lexicale
 - Apport d'expressivité par des variantes grapho-phonologiques
 - Phénomènes syntaxiques
 - Certains propres à l'écrit en temps réel
 - Peuvent se rapprocher de la langue parlée



Présentation du corpus

- Quantification:
 - 3 mois de discussions
 - 105 canaux publics, récréatifs et techniques
 - 4 192 033 messages
 - 23 011 876 mots
 - 5,5 mots / message



Présentation du corpus

- Grande variété de canaux
 - Sur le plan des utilisateurs, avec des canaux destinés à:
 - Une tranche d'âge: #18-25ans
 - Une localisation: #quebec
 - Un groupe: #hokutoteam
 - Sur le plan thématique
 - Discussion générale: #-p-u-r-, #darkcloud, #enjoy, #edelweiss
 - Discussion thématique:
 - Informatique: #debian
 - Programmation: #c++, #scripts
 - Actualité: #actu
 - Jeux vidéos: #allsoluces, #cstrike
 - Dragage: #endensensuelcalm, #caline
 - Sur le plan linguistique



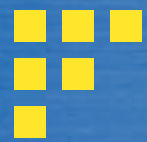
Présentation du corpus

- Collecte

- Sur le réseau IRC Epiknet, via le service BotStats
 - Service d'archivage de sessions de tchat
 - A l'initiative du propriétaire du canal
 - Publication des archives avec ou sans mots de passe

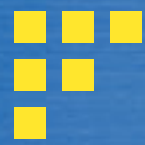
- Droit d'auteur

- Tchat = œuvre collective publiée sous pseudonyme
- L'éditeur est le dépositaire des droits d'auteur
- Editeur = service de publication (BotStats)
- Autorisation de reproduction et de diffusion



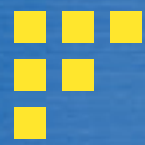
Structure du corpus

- Reformaté en XML
 - Pour permettre l'annotation et le traitement automatique

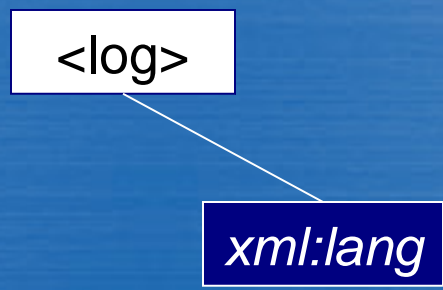


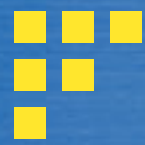
Structure du corpus

<log>

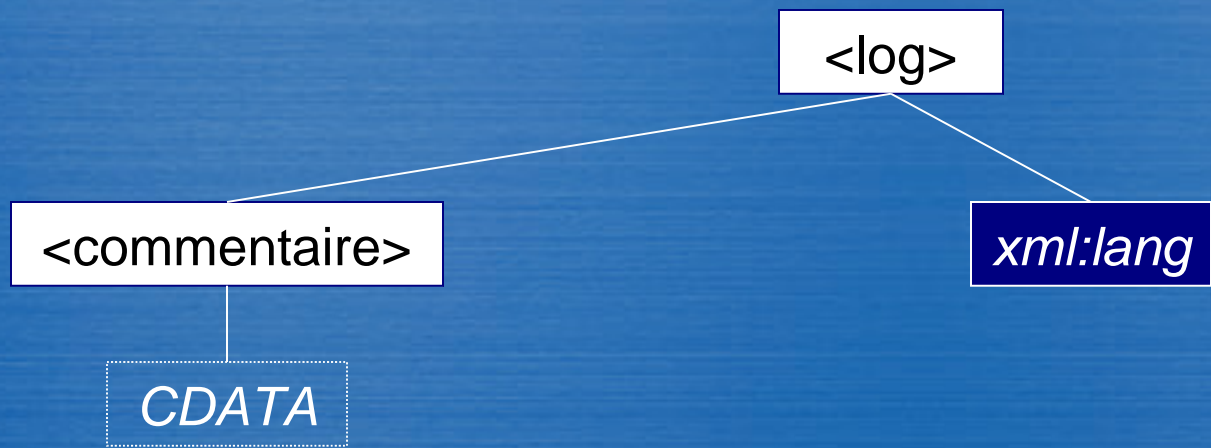


Structure du corpus



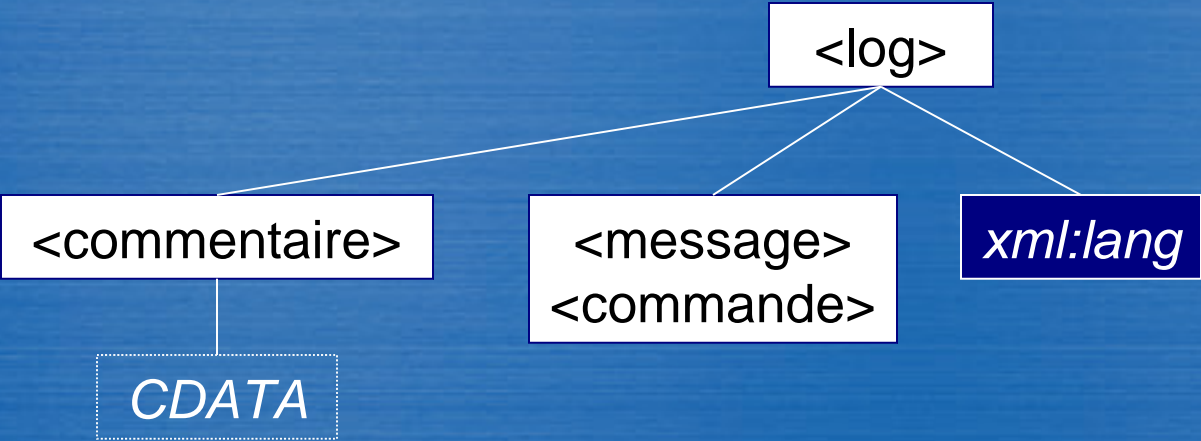


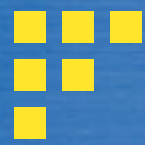
Structure du corpus



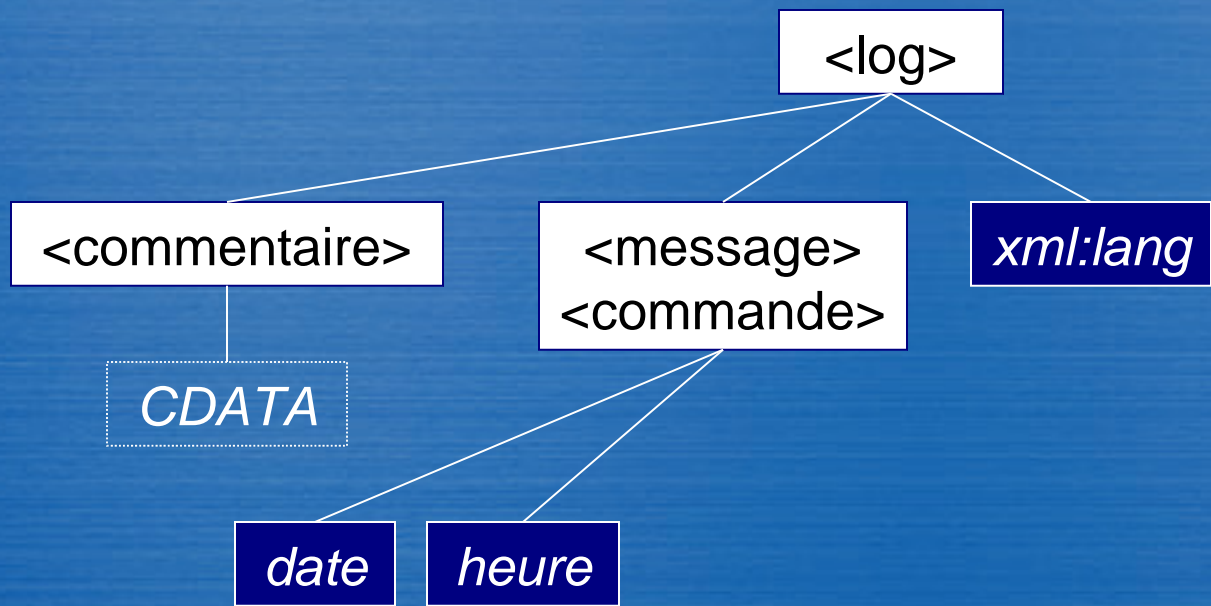


Structure du corpus

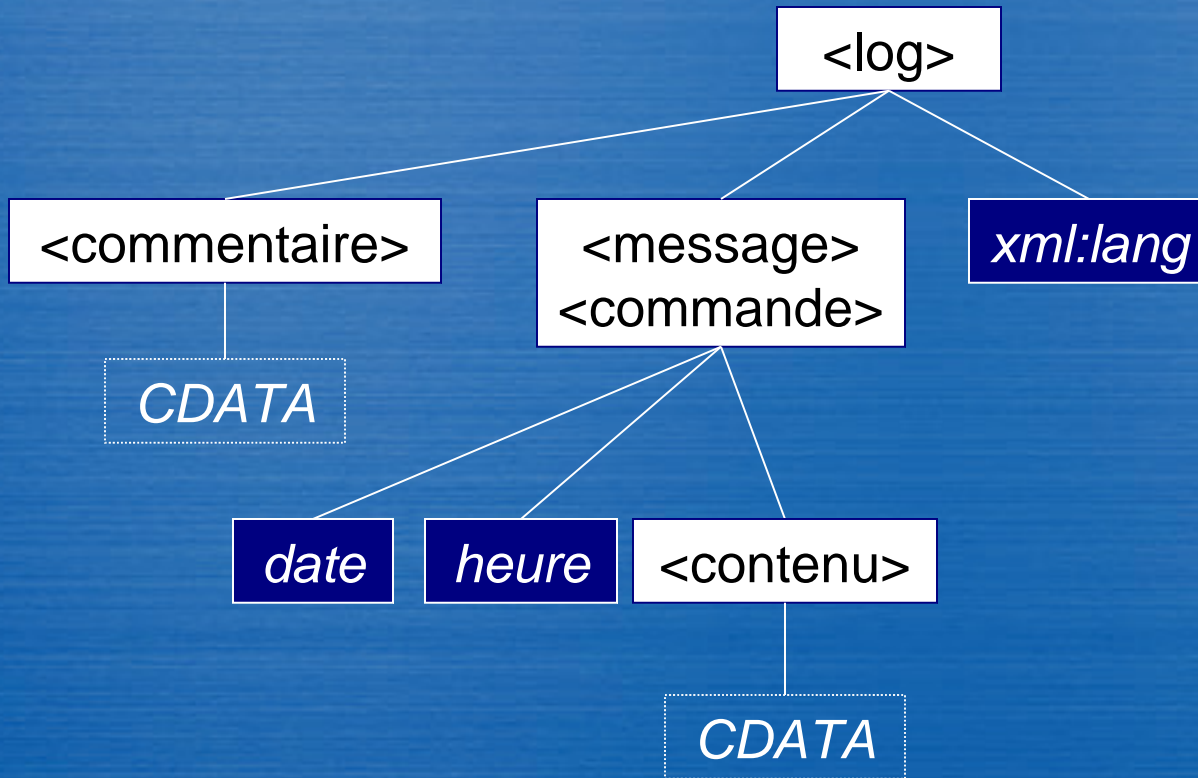




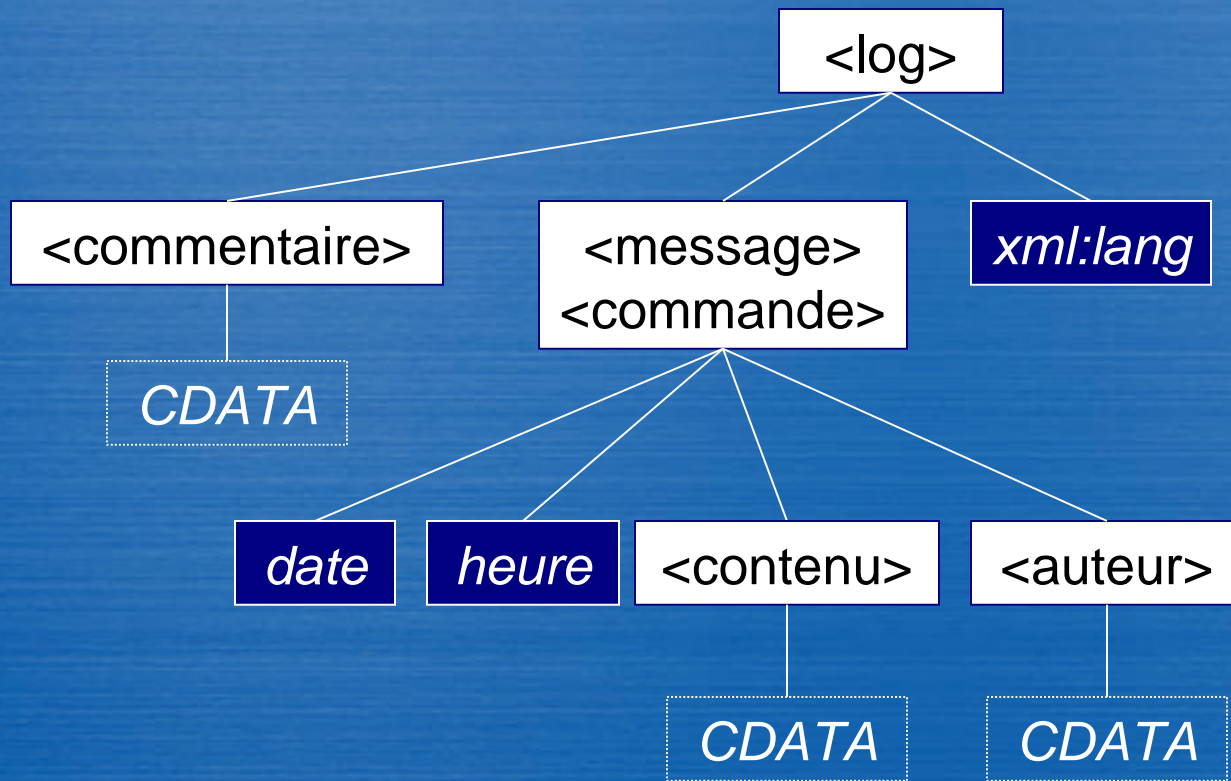
Structure du corpus



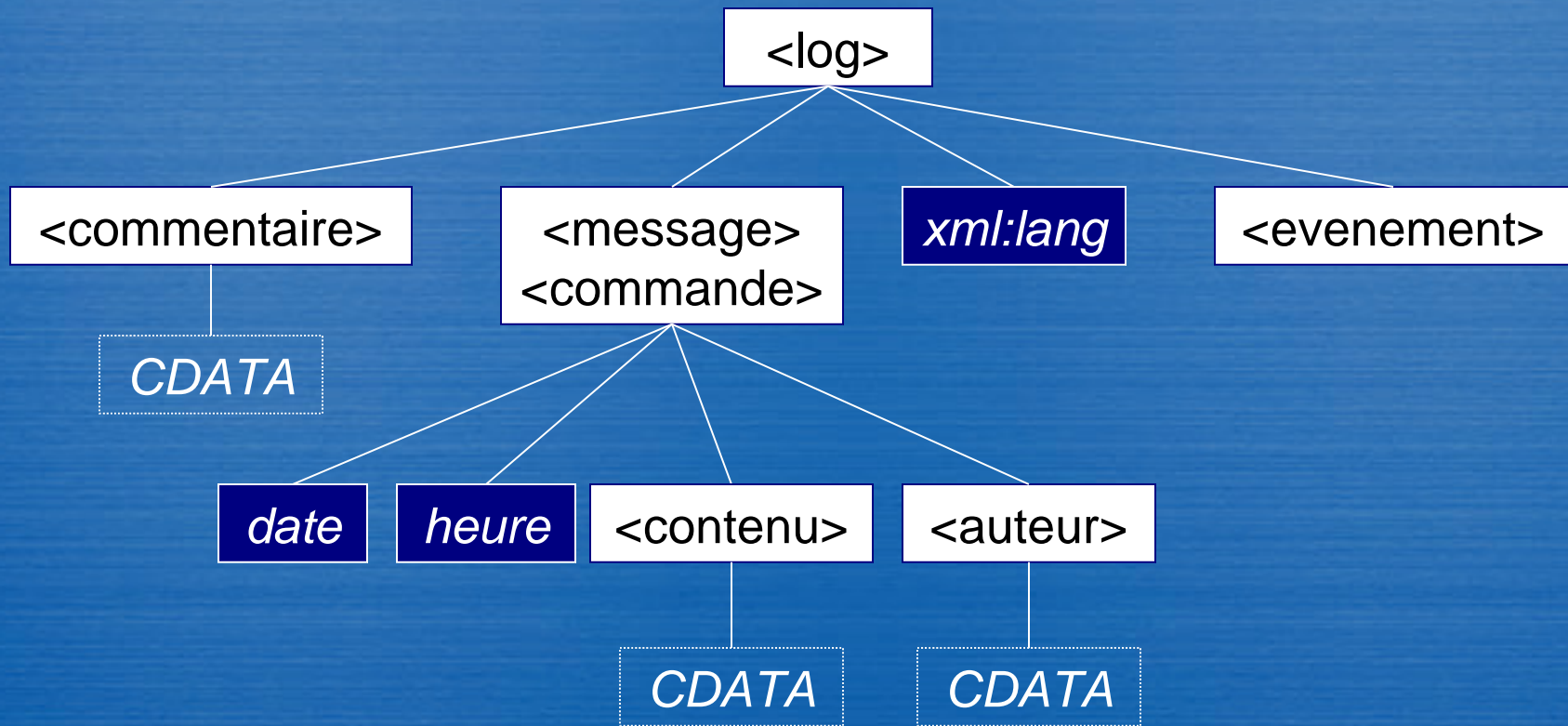
Structure du corpus



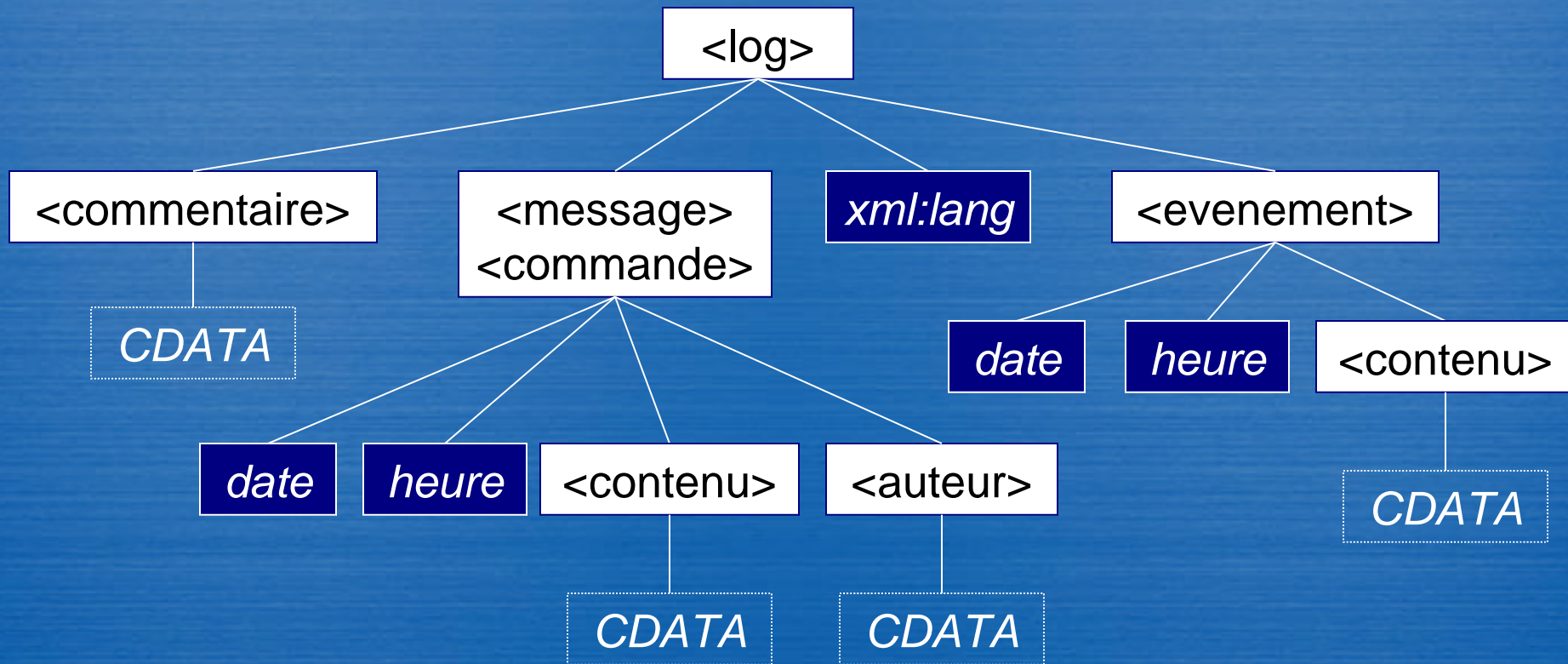
Structure du corpus



Structure du corpus



Structure du corpus



Publication du corpus

The screenshot shows a Mozilla Firefox browser window with the address bar at <http://www.clips.imag.fr/geta/User/achille.falaise/corpuschat/index.html>. The page title is "Corpus de français tchaté".

Canal

Corpus de français tchaté

Qu'est-ce que c'est ?

Ce site présente un **corpus de français tchaté** de 4 millions de messages, soit environ **23 millions de mots**, correspondant à 3 mois de conversations du printemps 2004, sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs disponibles sur le site <http://www.botstats.com>. Ce site publie les logs de quelques centaines de canaux du serveur IRC **EpikNet**. La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les **logs publics** de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est **formaté en XML**, et est associé à une **feuille de style XSL**. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veillez-donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus conviviale.

Navigateur requis

Un navigateur supportant XML/XSL:

- **Firefox 0.9+**: le navigateur open-source de la **fondation Mozilla** est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.
- **Internet Explorer 5+**: le navigateur de **Microsoft** supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la **librairie MSXML**. La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

Les canaux du corpus

Nom	Description	Messages	%
#--quizz--		57917	2.3
#p-u-r-		11811	0.47
#18-25ans		527290	20.9
#[dmb]dreamchan		67332	2.67
#actu		15582	0.62
#allsoluces		19481	0.77
#anaïsgirl		106284	4.21
#angel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c++		43399	1.72
#caline		34127	1.35
#cstrike		24628	0.98
#darkcloud		49913	1.98
#dbz_legend		50464	2
#debian		1040	0.04
#deglingo		4598	0.18
#dragon-ball-z		8465	0.34
#edelweiss		205446	8.14
#edensuelcam		199874	7.92
#enjoy		77621	3.08

Terminé

Retour à la présentation...

Canal
[dropdown menu]

Corpus de français tchaté

Qu'est-ce que c'est ?

Ce site présente un **corpus de français tchaté** de 4 millions de messages, soit environ **23 millions de mots**, correspondant à 3 mois de conversations du printemps 2004, sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs disponibles sur le site <http://www.botstats.com>. Ce site publie les logs de quelques centaines de canaux du serveur IRC **EpikNet**. La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les **logs publics** de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est **formaté en XML**, et est associé à une **feuille de style XSL**. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veuillez donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus conviviale.



Les canaux du corpus

Nom	Description	Messages	%
#--quizz--		57917	2.3
#-p-u-r-		11811	0.47
#18-25ans		527290	20.9
#[dmb]dreamchan		67332	2.67
#actu		15582	0.62
#allsoluces		19481	0.77
#anaigirl		106284	4.21
#angel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c++		43399	1.72
#caline		34127	1.35
#cstrike		24628	0.98
#darkcloud		49913	1.98
#dbz_legend		50464	2
#debian		1040	0.04
#deglingo		4598	0.18
#drag-robots		1445	0.06
#edelweiss		23446	0.94
#edensersuelcam		79874	3.12
#enjoy		1721	0.07

Navigateurs compatibles
(supportant les feuilles
de style XSL)

Navigateur requis

Un navigateur supportant XML/XSL:

-  - **Firefox 0.9+**: le navigateur open-source de la **fondation Mozilla** est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.
-  - **Internet Explorer 5+**: le navigateur de **Microsoft** supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la **librairie MSXML**. La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

☑ Firefox 0.9+

☑ Linux

☑ MacOS

☑ Windows

• ☑ Internet Explorer 4.5+

☒ MacOS

☑ Windows 98+ (requiert la librairie MSXML)

☑ Windows XP

☑ Mozilla 1.7+

☑ Linux

☑ Windows

☒ Opera

☒ Linux

☒ MacOS

☒ Windows

☑ Safari 1.3+ (MacOS)

Corpus de français tchaté

Qu'est-ce que c'est ?

Le site présente un corpus de français tchaté de 4 millions de messages, soit environ 23 millions de mots, correspondant à 3 mois de conversations du printemps 2004, sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs de quelques centaines de canaux du serveur IRC EpikNet. La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les logs publics de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est formaté en XML, et est associé à une feuille de style XSL. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veuillez donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus

Les canaux du corpus

Non	Description	Messages	%
#--qizz--		57917	2.3
#p-r		11811	0.47
#1825ans		527290	20.9
#[dr]b]dreamchan		67332	2.67
#actu		15582	0.62
#alloluces		19481	0.77
#andisgirl		106284	4.21
#andiel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c+		43399	1.72
#caline		34127	1.35
#castrike		24628	0.98
#darkcloud		49913	1.98
#dblegend		50464	2
#delian		1040	0.04
#delingo		4598	0.18
#delig-1-ans		449	0.02
#elweiss		23446	0.14
#elinsersuelcam		79874	7.12
#eloy		771	0.03

Navigateurs compatibles (supportant les feuilles de style XSL)

Navigateur requis

Un navigateur supportant XML/XSL:

- **Firefox 0.9+**: le navigateur open-source de la fondation Mozilla est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.

- **Internet Explorer 5+**: le navigateur de Microsoft supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la librairie MSXML. La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

Retour à la présentation...

Retour à la page perso...

Canal

Corpus de français tchaté

Sélection du canal



Ce site présente un **corpus de français tchaté** de 4 millions de messages, soit environ **23 millions de mots**, correspondant à 3 mois de conversations du printemps 2004, sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs disponibles sur le site <http://www.botstats.com>. Ce site publie les logs de quelques centaines de canaux du serveur IRC [EpikNet](#). La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les **logs publics** de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est **formaté en XML**, et est associé à une **feuille de style XSL**. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veuillez-donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus conviviale.

Navigateur requis

Un navigateur supportant XML/XSL:

-  - **Firefox 0.9+**: le navigateur open-source de la [fondation Mozilla](#) est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.
-  - **Internet Explorer 5+**: le navigateur de [Microsoft](#) supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la [bibliothèque MSXML](#). La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

Les canaux du corpus

Nom	Description	Messages	%
#--quizz--		57917	2.3
#p-u-r-		11811	0.47
#18-25ans		527290	20.9
#[dmb]dreamchan		67332	2.67
#actu		15582	0.62
#allsoluces		19481	0.77
#anaigirl		106284	4.21
#angel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c++		43399	1.72
#caline		34127	1.35
#cstrike		24628	0.98
#darkcloud		49913	1.98
#dbz_legend		50464	2
#debian		1040	0.04
#deglingo		4598	0.18
#dragon-ball-z		8465	0.34
#edelweiss		205446	8.14
#edensuelcam		199874	7.92
#enjoy		77621	3.08

Canal

18-25ans

Période

Date de début:

Date de fin:

Messages/page

100

Valider

Corpus de français tchaté

Sélection de la période et du nombre de messages par page



Ce site présente un **corpus de français tchaté** de 4 millions de messages provenant de 105 canaux de chats correspondant à 3 mois de conversations du printemps 2004 sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs disponibles sur le site <http://www.botstats.com>. Ce site publie les logs de quelques centaines de canaux du serveur IRC **EpikNet**. La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les **logs publics** de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est **formaté en XML**, et est associé à une **feuille de style XSL**. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veuillez-donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus conviviale.

Navigateur requis

Un navigateur supportant XML/XSL:

-  - **Firefox 0.9+**: le navigateur open-source de la [fondation Mozilla](http://www.mozilla.org) est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.
-  - **Internet Explorer 5+**: le navigateur de [Microsoft](http://www.microsoft.com) supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la [librairie MSXML](#). La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

Les canaux du corpus

Nom	Description	Messages	%
#--quizz--		57917	2.3
#p-u-r-		11811	0.47
#18-25ans		527290	20.9
#[dmb]dreamchan		67332	2.67
#actu		15582	0.62
#allsoluces		19481	0.77
#anaïsgirl		106284	4.21
#angel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c++		43399	1.72
#caline		34127	1.35
#cstrike		24628	0.98
#darkcloud		49913	1.98
#dbz_legend		50464	2
#debian		1040	0.04
#deglingo		4598	0.18
#dragon-ball-z		8465	0.34
#edelweiss		205446	8.14
#edensuelcam		199874	7.92
#enjoy		77621	3.08

Retour à la présentation...

Corpus de français tchaté

Retour à la page perso...

Canal

18-25ans

Période

Date de début:

10/01/2004

Date de fin:

20/01/2004

Messages/page

2000

Valider

Qu'est-ce que c'est ?

Ce site présente un **corpus de français tchaté** de 4 millions de messages, soit environ **23 millions de mots**, correspondant à 3 mois de conversations du printemps 2004, sur 105 canaux différents. Le corpus de tchat est constitué à partir des logs disponibles sur le site <http://www.botstats.com>. Ce site publie les logs de quelques centaines de canaux du serveur IRC [EpikNet](http://www.epiknet.com). La publication de ces logs est un choix de la part du créateur du canal; ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux ont décidé de les rendre accessibles à tous, et ce sont les **logs publics** de ces canaux qui sont regroupés au sein du corpus.

Comment ça marche ?

Le menu situé à gauche de cette fenêtre permet de naviguer dans le corpus. Celui-ci est **formaté en XML**, et est associé à une **feuille de style XSL**. Cela signifie que le code XML est toujours accessible à l'aide de la fonction "afficher le code source" de votre navigateur (attention, ce site contient des cadres, veuillez-donc à appliquer cette fonction au cadre dans lequel le corpus est affiché !). Mais le code XML n'étant pas nécessairement très parlant pour tout le monde, la feuille de style XSL permet de le visualiser sous une forme plus conviviale.

Navigateur requis

Un navigateur supportant XML/XSL:

- **Firefox 0.9+**: le navigateur open-source de la [fondation Mozilla](http://www.mozilla.org) est disponible gratuitement pour les systèmes d'exploitation les plus courants (Linux, Mac OS X et Windows). Ce site vise avant tout la compatibilité avec ce navigateur, et son utilisation est donc vivement conseillée.
- **Internet Explorer 5+**: le navigateur de [Microsoft](http://www.microsoft.com) supporte XML/XSL **uniquement avec le système d'exploitation Windows**. Pour les systèmes d'exploitation antérieurs à Windows XP, il peut être nécessaire de télécharger la [librairie MSXML](#). La compatibilité de ce site avec Internet Explorer n'est pas complète, mais celui-ci reste utilisable.

Les canaux du corpus

Nom	Description	Messages	%
#--quizz--		57917	2.3
#p-u-r-		11811	0.47
#18-25ans		527290	20.9
#[dmb]dreamchan		67332	2.67
#actu		15582	0.62
#allsoluces		19481	0.77
#anaïsgirl		106284	4.21
#angel-corp		116573	4.62
#blondin		14630	0.58
#botstats		18143	0.72
#c++		43399	1.72
#caline		34127	1.35
#cstrike		24628	0.98
#darkcloud		49913	1.98
#dbz_legend		50464	2
#debian		1040	0.04
#deglingo		4598	0.18
#dragon-ball-z		8465	0.34
#edelweiss		205446	8.14
#edensensuelcam		199874	7.92
#enjoy		77621	3.08

Vue du corpus

Retour à la présentation...	Num	Date	Heure	Auteur	Texte
Canal 18-25ans	22003	10/01/2004	00:00	DataStats02	::: Les statistiques de ce canal sont disponibles sur: http://go.botstats.com/?18-25
Période Date de début: 10/01/2004 Date de fin: 20/01/2004	22004	10/01/2004	00:00	Changement de pseudo: B2 -> Ragondin	
Messages/page 2000	22005	10/01/2004	00:00	Rainner .{AwaY}. .{suiko powa}.	
Valider	22006	10/01/2004	00:00	Changement de pseudo: Rainner -> Rainner_AwAy	
Page de résultats 1 ->	22007	10/01/2004	00:00	Shaolan	lut Paradise
	22008	10/01/2004	00:00	Shaolan	lut Paris
	22009	10/01/2004	00:01	Paris	hello
	22010	10/01/2004	00:01	Paris	ca va le peumle???????
	22011	10/01/2004	00:01	Paris	le peuple
	22012	10/01/2004	00:01	Melina	ui
	22013	10/01/2004	00:02	D-espace is listening to Pixies - Monkey Gone to Heaven ` R-script ,	
	22014	10/01/2004	00:02	Lou- (~Anonyme@EppLqTgQKshW.) vient d'arriver ! #18-25ans.	
	22015	10/01/2004	00:02	Shaolan	lut famas
	22016	10/01/2004	00:02	Lou-	re
	22017	10/01/2004	00:02	Shaolan	re Lou-
	22018	10/01/2004	00:02	Changement de pseudo: famas -> Anonyme3260203	
	22019	10/01/2004	00:02	Lou-	vous n'avez pas vu paradise ?
	22020	10/01/2004	00:02	Lou-	il vient de quitter ?
	22021	10/01/2004	00:02	(~pogman_h4@Epojvk6Hprb.g) s'est déconnecté: Client exited	
	22022	10/01/2004	00:02	Paris	I m here
	22023	10/01/2004	00:03	Paris	T
	22024	10/01/2004	00:03	Paris	R
	22025	10/01/2004	00:04	Paris	I
	22026	10/01/2004	00:04	D-espace is listening to Bob Marley - Buffalo Soldier ` R-script ,	
	22027	10/01/2004	00:05	(~famas@EpQfWi.BO4O3o) s'est déconnecté: Client exited	
	22028	10/01/2004	00:05	Lou-	lol
	22029	10/01/2004	00:05	Paris	je reviens
	22030	10/01/2004	00:05	Paradise	guess who s back??????
	22031	10/01/2004	00:05	(DeMoN@Epcu5h/BfaIkA) vient d'arriver ! #18-25ans.	

Retour à la présentation...

Canal

18-25ans

Période

Date de début:

10/01/2004

Date de fin:

20/01/2004

Messages/page

2000

Valider

Page de résultats

1

->

Sélection de la
page de résultats

Num	Date	Heure	Auteur	Texte
22003	10/01/2004	00:00	DataStats02	::: Les statistiques de ce canal sont disponibles sur: http://go.botstats.com/?18-25ans
22004	10/01/2004	00:00	Changement de pseudo: B2 -> Ragondin	
22005	10/01/2004	00:00	Rainner .{AwAY}. .{suiko powa}.	
22006	10/01/2004	00:00	Changement de pseudo: Rainner -> Rainner_AwAy	
22007	10/01/2004	00:00	Shaolan	lut Paradise
22008	10/01/2004	00:00	Shaolan	lut Paris
22009	10/01/2004	00:01	Paris	hello
22010	10/01/2004	00:01	Paris	ca va le peumle???????
22011	10/01/2004	00:01	Paris	le peuple
22012	10/01/2004	00:01	Melina	ui
22013	10/01/2004	00:02	D-espace is listening to Pixies - Monkey Gone to Heaven ` R-script ,	
22014	10/01/2004	00:02	Lou- (~Anonyme@EppLqTgQKshW.) vient d'arriver ! #18-25ans.	
22015	10/01/2004	00:02	Shaolan	lut famas
22016	10/01/2004	00:02	Lou-	re
22017	10/01/2004	00:02	Shaolan	re Lou-
22018	10/01/2004	00:02	Changement de pseudo: famas -> Anonyme3260203	
22019	10/01/2004	00:02	Lou-	vous n'avez pas vu paradise ?
22020	10/01/2004	00:02	Lou-	il vient de quitter ?
22021	10/01/2004	00:02	(~pogman_h4@Epojvk6Hprb.g) s'est déconnecté: Client exited	
22022	10/01/2004	00:02	Paris	I m here
22023	10/01/2004	00:03	Paris	T
22024	10/01/2004	00:03	Paris	R
22025	10/01/2004	00:04	Paris	I
22026	10/01/2004	00:04	D-espace is listening to Bob Marley - Buffalo Soldier ` R-script ,	
22027	10/01/2004	00:05	(~famas@EpQfWi.BO4O3o) s'est déconnecté: Client exited	
22028	10/01/2004	00:05	Lou-	lol
22029	10/01/2004	00:05	Paris	je reviens
22030	10/01/2004	00:05	Paradise	guess who s back??????
22031	10/01/2004	00:05	(DeMoN@Epcu5h/BfaIkA) vient d'arriver ! #18-25ans.	



Retour à la présentation...

Canal

18-25ans

Période

Date de début:

10/01/2004

Date de fin:

20/01/2004

Messages/page

2000

Valider

Page de résultats

<- 9 ->

Num	Date	Heure	Auteur	Texte
38003	12/01/2004	17:31	TripI-MaKiNa	Hello people
38004	12/01/2004	17:31	Romeo^^	leclown wai mais chui pas homo
38005	12/01/2004	17:32	Shamanking	bonsoir tout le monde
38006	12/01/2004	17:32	leclown	!place
38007	12/01/2004	17:32	leclown	mouarf
38008	12/01/2004	17:32	leclown	lol
38009	12/01/2004	17:32	Avrilav	izzo: ils cherchent un chanteur à nantes à l'anpe, tu veux pas ? LOL
38010	12/01/2004	17:32	leclown	ouaou 51 cool
38011	12/01/2004	17:32	Romeo^^	leclown jimage chii avec des cheveux cours pour toi mdr
38012	12/01/2004	17:32	leclown	7:31 Notice DataStats02 leclown se classe en #51 position sur 909.
38013	12/01/2004	17:32	Romeo^^	=)
38014	12/01/2004	17:32	izzo	un chanteur de koi ?
38015	12/01/2004	17:32	leclown	mon numero preferer
38016	12/01/2004	17:32	leclown	lol
38017	12/01/2004	17:32	leclown	VIVE LE PASTIS
38018	12/01/2004	17:32	leclown	:)
38019	12/01/2004	17:32	Avrilav	CREATION D'UN CABARET DEBUT AVRIL 04
38020	12/01/2004	17:32	Avrilav	AVEC SPECTACLES VENDREDI ET SAMEDI
38021	12/01/2004	17:32	(~Kenshin@Ep/Wo.vA8pDyw)	vient d'arriver ! #18-25ans.
38022	12/01/2004	17:32	izzo	lol
38023	12/01/2004	17:32	Avrilav	SOIR + QUELQUES REPRESENTATIONS LE
38024	12/01/2004	17:32	Avrilav	DIMANCHE MIDI + POSSIBILITES DE
38025	12/01/2004	17:33	TrIpI-MaKiNa	VIVE LA MAKINA
38026	12/01/2004	17:33	izzo	euh
38027	12/01/2004	17:33	Avrilav	SPECTACLES EXCEPTIONNELS EN SEMAINE
38028	12/01/2004	17:33	Avrilav	- SI POSSIBLE PIANISTE -
38029	12/01/2004	17:33	Avrilav	CONNAISSANCES : SPECTACLE DE CABARET
38030	12/01/2004	17:33	cosette_lycee	vive le whisky!!
38031	12/01/2004	17:33	leclown	Avrilav arrete sinon Shaolan te mets dans une boite de conserve

Messages

Retour à la présentation...		Num	Date	Heure	Auteur	Texte
Canal		38003	12/01/2004	17:31	TripI-MaKiNa	Hello people
18-25ans		38004	12/01/2004	17:31	Romeo^^	leclown wai mais chui pas homo
Période		38005	12/01/2004	17:32	Shamanking	bonsoir tout le monde
Date de début:		38006	12/01/2004	17:32	leclown	!place
Date de fin:		38007	12/01/2004	17:32	leclown	mouarf
Messages/page		38008	12/01/2004	17:32	leclown	lol
2000		38009	12/01/2004	17:32	Avrilav	izzo: ils cherchent un chanteur à nantes à l'anpe, tu veux pas ? LOL
Valider		38010	12/01/2004	17:32	leclown	ouaou 51 cool
Page de résultats		38011	12/01/2004	17:32	Romeo^^	leclown jimage chii avec des cheveux courts pour toi mdr
<- 9 ->		38012	12/01/2004	17:32	leclown	7:31 Notice DataStats02 leclown se classe en #51 position sur 909.
		38013	12/01/2004	17:32	Romeo^^	=)
		38014	12/01/2004	17:32	izzo	un chanteur de koi ?
		38015	12/01/2004	17:32	leclown	mon numero preferer
		38016	12/01/2004	17:32	leclown	lol
		38017	12/01/2004	17:32	leclown	VIVE LE PASTIS
		38018	12/01/2004	17:32	leclown	:)
		38019	12/01/2004	17:32	Avrilav	CREATION D'UN CABARET DEBUT AVRIL 04
		38020	12/01/2004	17:32	Avrilav	AVEC SPECTACLES VENDREDI ET SAMEDI
		38021	12/01/2004	17:32	(~Kenshin@Ep/Wo.vA8pDyw)	vient d'arriver ! #18-25ans.
		38022	12/01/2004	17:32	izzo	lol
		38023	12/01/2004	17:32	Avrilav	SOIR + QUELQUES REPRESENTATIONS LE
		38024	12/01/2004	17:32	Avrilav	DIMANCHE MIDI + POSSIBILITES DE
		38025	12/01/2004	17:33	TrIpI-MaKiNa	VIVE LA MAKINA
		38026	12/01/2004	17:33	izzo	euh
		38027	12/01/2004	17:33	Avrilav	SPECTACLES EXCEPTIONNELS EN SEMAINE
		38028	12/01/2004	17:33	Avrilav	- SI POSSIBLE PIANISTE -
		38029	12/01/2004	17:33	Avrilav	CONNAISSANCES : SPECTACLE DE CABARET
		38030	12/01/2004	17:33	cosette_lycee	vive le whisky!!
		38031	12/01/2004	17:33	leclown	Avrilav arrete sinon Shaolan te mets dans une boite de conserve

Commande

Retour à la présentation...	Num	Date	Heure	Auteur	Texte
	38003	12/01/2004	17:31	TripI-MaKiNa	Hello people
	38004	12/01/2004	17:31	Romeo^^	leclown wai mais chui pas homo
	38005	12/01/2004	17:32	Shamanking	bonsoir tout le monde
	38006	12/01/2004	17:32	leclown	!place
	38007	12/01/2004	17:32	leclown	mouarf
	38008	12/01/2004	17:32	leclown	lol
	38009	12/01/2004	17:32	Avrilav	izzo: ils cherchent un chanteur à nantes à l'anpe, tu veux pas ? LOL
	38010	12/01/2004	17:32	leclown	ouaou 51 cool
	38011	12/01/2004	17:32	Romeo^^	leclown jimage chii avec des cheveux courts pour toi mdr
	38012	12/01/2004	17:32	leclown	7:31 Notice DataStats02 leclown se classe en #51 position sur 909.
	38013	12/01/2004	17:32	Romeo^^	=)
	38014	12/01/2004	17:32	izzo	un chanteur de koi ?
	38015	12/01/2004	17:32	leclown	mon numero preferer
	38016	12/01/2004	17:32	leclown	lol
	38017	12/01/2004	17:32	leclown	VIVE LE PASTIS
	38018	12/01/2004	17:32	leclown	:)
	38019	12/01/2004	17:32	Avrilav	CREATION D'UN CABARET DEBUT AVRIL 04
	38020	12/01/2004	17:32	Avrilav	AVEC SPECTACLES VENDREDI ET SAMEDI
	38021	12/01/2004	17:32	(~Kenshin@Ep/Wo.vA8pDyw)	vient d'arriver ! #18-25ans.
	38022	12/01/2004	17:32	izzo	lol
	38023	12/01/2004	17:32	Avrilav	SOIR + QUELQUES REPRESENTATIONS LE
	38024	12/01/2004	17:32	Avrilav	DIMANCHE MIDI + POSSIBILITES DE
	38025	12/01/2004	17:33	TripI-MaKiNa	VIVE LA MAKINA
	38026	12/01/2004	17:33	izzo	euh
	38027	12/01/2004	17:33	Avrilav	SPECTACLES EXCEPTIONNELS EN SEMAINE
	38028	12/01/2004	17:33	Avrilav	- SI POSSIBLE PIANISTE -
	38029	12/01/2004	17:33	Avrilav	CONNAISSANCES : SPECTACLE DE CABARET
	38030	12/01/2004	17:33	cosette_lycee	vive le whisky!!
	38031	12/01/2004	17:33	leclown	Avrilav arrete sinon Shaolan te mets dans une boite de conserve

Evénement

Retour à la présentation...

Canal
18-25ans

Période
Date de début: 10/01/2004
Date de fin: 20/01/2004

Messages/page
2000

Valider

Page de résultats
9

Num	Date	Heure	Auteur	Texte
38003	12/01/2004	17:31	TriPi-MaKiNa	Hello people
38004	12/01/2004	17:31	Romeo^^	leclown wai mais chui pas homo
38005	12/01/2004	17:32	Shamanking	bonsoir tout le monde
38006	12/01/2004	17:32	leclown	Inlace

view-source: - Source de : http://www-clips.imag.fr/geta/User/achille.falaise/corpuschat/corpus.php3?lin

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet type="text/xsl" href="corpus.xsl"?>
<log xml:lang="fr"><commentaire>Canal #18-25ans</commentaire><message num="38003" jour="12" mois="01" a
<message num="38004" jour="12" mois="01" annee="2004" heure="17" min="31"><auteur>Romeo^^</auteur><cont
<message num="38005" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Shamanking</auteur><c
<commande num="38006" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cor
<message num="38007" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38008" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38009" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Avrilav</auteur><cont
<message num="38010" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38011" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Romeo^^</auteur><cont
<message num="38012" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38013" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Romeo^^</auteur><cont
<message num="38014" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><conten
<message num="38015" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38016" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
<message num="38017" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
38021 12 <message num="38018" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>leclown</auteur><cont
38022 12 <message num="38019" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Avrilav</auteur><cont
<message num="38020" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Avrilav</auteur><cont
38023 12 <evenement num="38021" jour="12" mois="01" annee="2004" heure="17" min="32"><contenu>(-Kenshin@Ep/Wo.vA
38024 12 <message num="38022" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>izzo</auteur><conten
<message num="38023" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Avrilav</auteur><cont
<message num="38024" jour="12" mois="01" annee="2004" heure="17" min="32"><auteur>Avrilav</auteur><cont
38026 12 <message num="38025" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>TriPi-MaKiNa</auteur>
<message num="38026" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>izzo</auteur><conten
<message num="38027" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>Avrilav</auteur><cont
38028 12 <message num="38028" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>Avrilav</auteur><cont
38029 12 <message num="38029" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>Avrilav</auteur><cont
38030 12 <message num="38030" jour="12" mois="01" annee="2004" heure="17" min="33"><auteur>cosette_lycee</auteur>
38031 12/01/2004 17:33 leclown Avrilav arrete sinon sinon le mets dans une boîte de conserve

```

Corpus au format XML



Conclusion et perspectives

- Résultats:
 - Plus important corpus de tchat disponible à l'heure actuelle
 - Version démo disponible sur:
<http://www-clips.imag.fr/geta/User/achille.falaise/corpuستchat/>
- Utilisation d'un corpus de tchat:
 - Pour l'étude:
 - Du dialogue
 - Des communautés
 - De la personnalisation et de l'oralisation de l'écrit
 - Des interactions avec des agents informatiques
 - Etc...
 - Pour l'élaboration de modèles de langage pour la reco
 - Plus adapté qu'un corpus issu du web ou d'articles de presse



Conclusion et perspectives

- Projets:
 - Corpus monolingue
 - Ajout de quelques nouveaux outils à l'interface
 - Corpus multilingue
 - Thèse sur la TA de dialogues oraux
 - Mise en place d'un système de tchat pour utilisateur imparfaitement bilingue, avec « sous-titrage » partiel en TA
 - Collecte d'un corpus multilingue
 - Evaluation de l'utilisabilité d'une telle « béquille »

RECITAL 2005

Constitution d'un corpus de français tchaté

<http://www-clips.imag.fr/geta/User/achille.falaise/corpuستchat/>

Questions...

Remerciements:



GETA
CLIPS

BotStats

Epiknet

Prosodie