Alpes4science : corpus de SMS réels dans les Alpes

Georges Antoniadis, Virginie Zampa
Laboratoire LIDILEM, Université Stendhal de Grenoble
Georges.Antoniadis@u-grenoble3.fr, Virginie.Zampa@u-grenoble3.fr

Groupe de travail Nouv-com, IR corpus-écrits, séminaire en ligne du 31 mai 2012

Alpes4science : Buts et préliminaires

A l'origine : Centre de l'Oralité Alpine, Département des Hautes-Alpes, Marc Mallen

L'intermédiaire : Laboratoire CENTAL, Université Catholique de Louvain, Belgique, projet sms4science

Collecte : Principalement en Isère et Hautes-Alpes

But du CG05 : Constituer un corpus de sms réels, outil pour l'étude de « l'oralité » dans les Alpes

Alpes4science : Buts et préliminaires

Buts du LIDILEM

Constituer un corpus de sms réels

- Premier corpus de ce type en France métropolitaine
- Obtenir des données sociolinguistiques des utilisateurs de sms (questionnaire en ligne)
- Mettre à disposition, gratuitement, le corpus pour des fins de recherche

Lancer des travaux de recherche sur les données recueillies

- Etude du « langage » sms
- Applications liées à l'utilisation de sms

Lancer des études comparatives avec d'autres corpus de sms francophones (Québec, La Réunion, Belgique), voire d'autres langues

Alpes4science : La collecte

A la recherche d'un opérateur téléphonique

- Les opérateurs français ne sont pas intéressés par le projet
- Les numéros vert pour les portables n'existent pas en France, coût pour les donneurs
- Il faut payer aussi pour chaque sms reçu !
- Solutions payantes possibles via entreprises spécialisés ou opérateurs téléphoniques

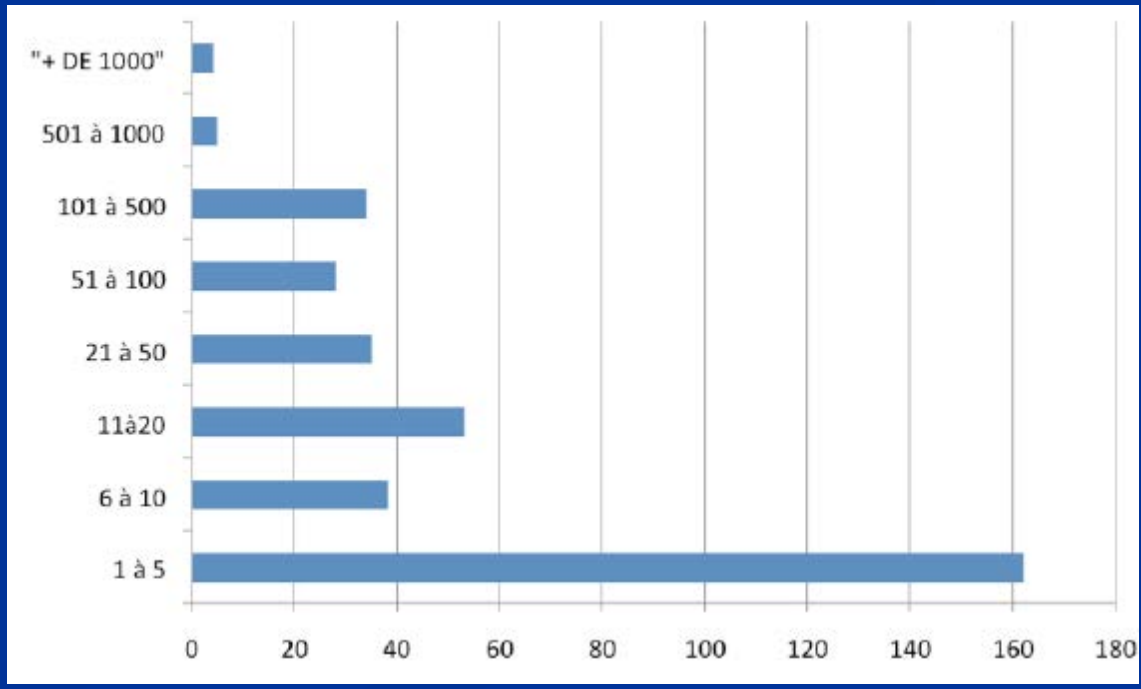
- Plateforme ORANGE grâce au CG05

LIDILEM s'associe à la collecte et prend en charge la plus grosse partie

- Campagne de communication, politique de cadeaux (~0,25€/sms reçu)

Alpes4science : La collecte

- Quatre mois de collecte
- 22 054 SMS collectés
- 359 participants
- 240 questionnaires liés à des participants
- 170 femmes et 70 hommes
- 84 utilisateurs de clavier azerty



Alpes4science : La collecte

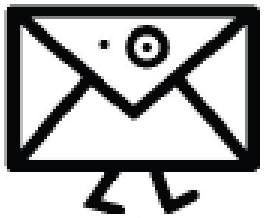
sms05 + SMS

31014

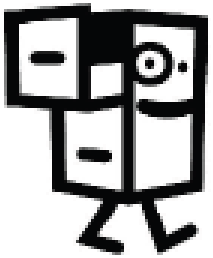
Numéro court

orange

Plateforme
"Contact Everyone"



Notifications par mail



Base de données

Problèmes : SMS longs, SMS Unicode, SMS perdus…

Alpes4science : Après la collecte

- Anonymisation des SMS
- Transcription des SMS
- Dictionnaire à double entrée
- Diverses analyses sur les données

Alpes4science : Anonymisation

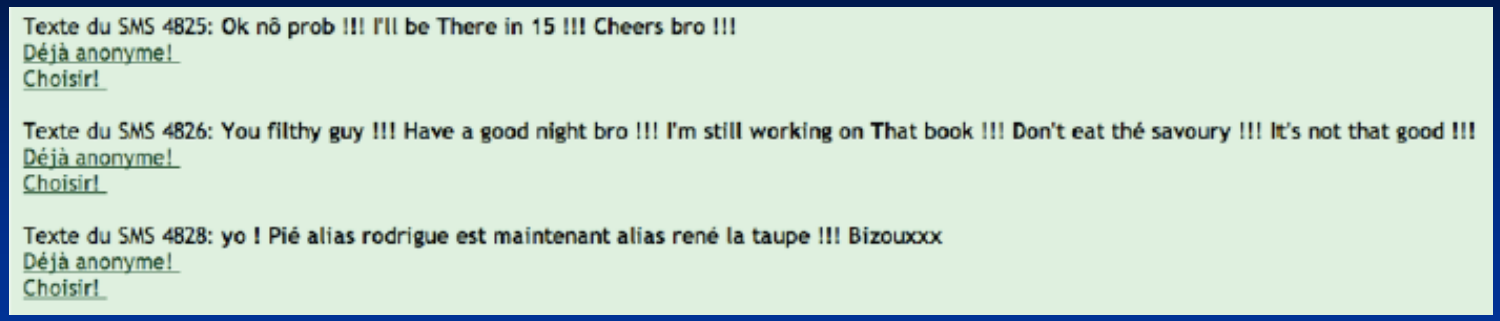
Anonymisation… non automatique

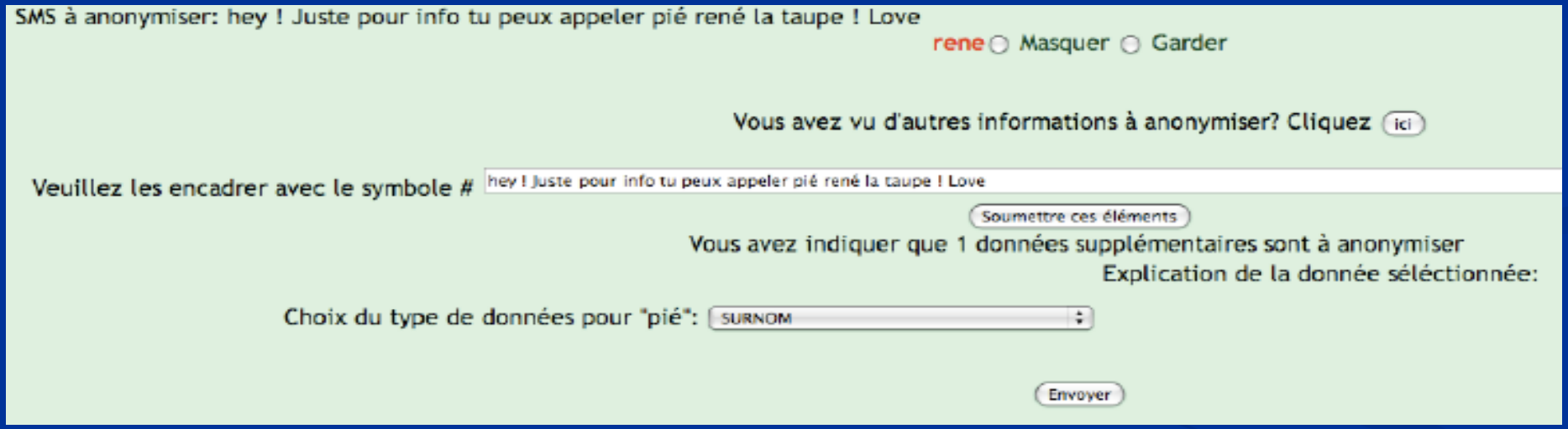
- anonymisation assistée
- table des prénoms
- reconnaissance de patern (code bancaires,
- telephone, etc.)

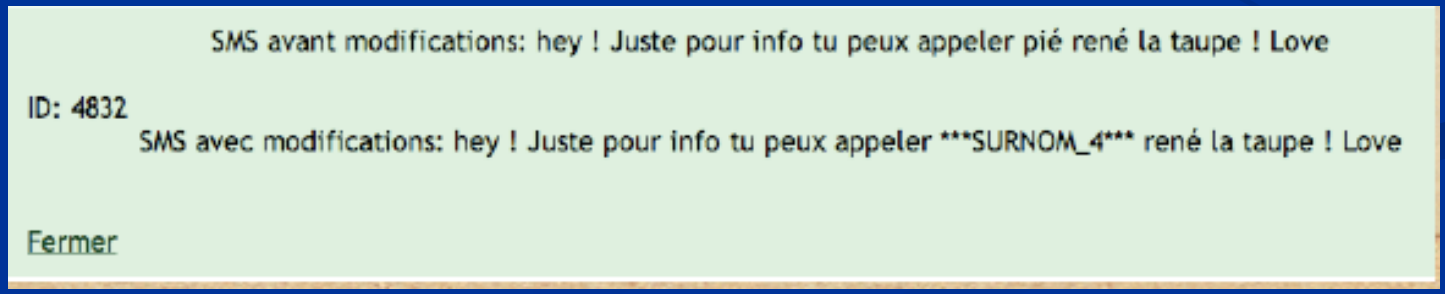
Pourquoi non automatique ?

- TL serait parfait en Arthur en plus non ? Et, OMG, je viens d'y penses, plutôt SG en Guenièvre ;-)

Alpes4science : Interface d'anonymisation



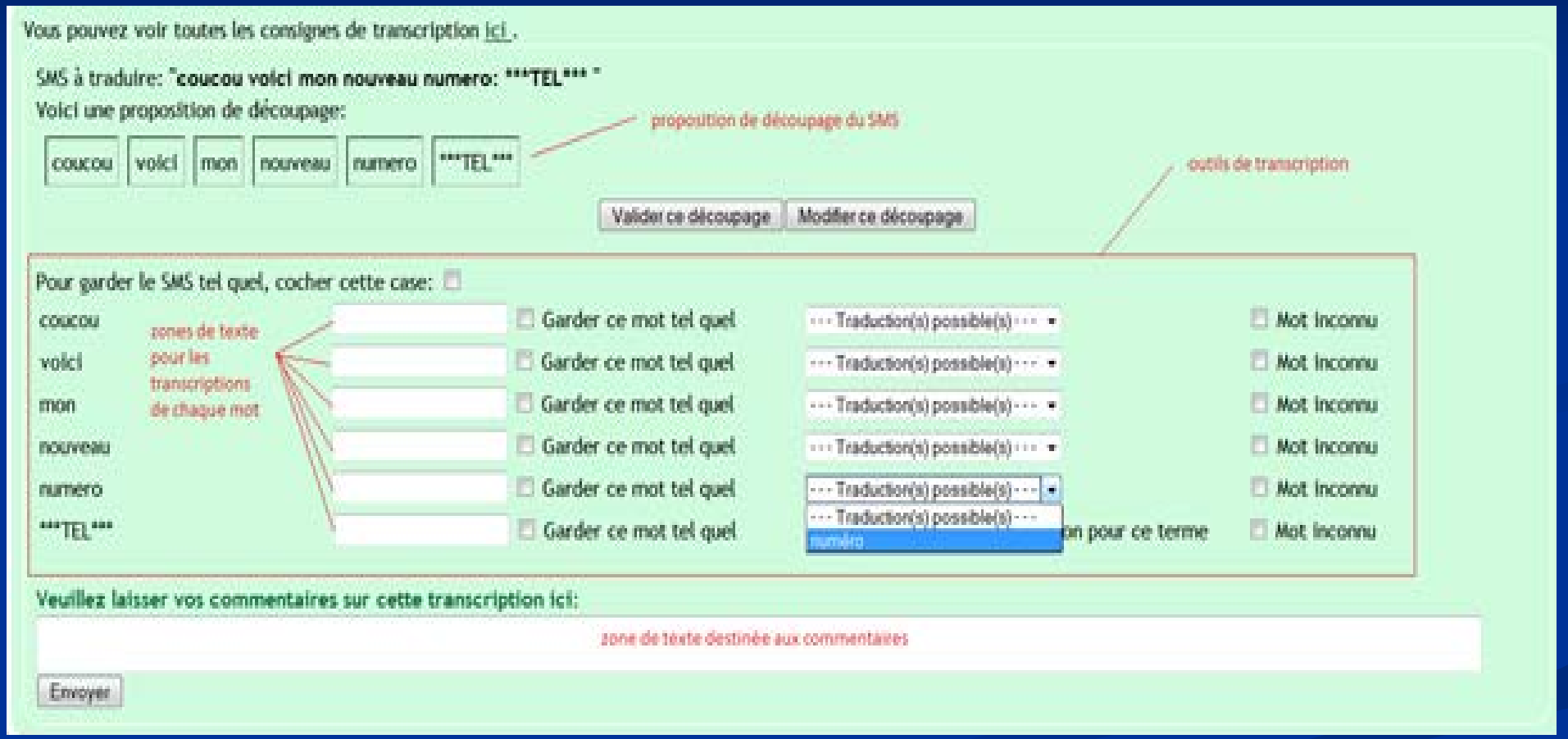




Alpes4science : Transcription

- Fautes : corrigées
- Mots autre langue : corrigés si nécessaire + traduction en commentaire
- Smileys : préservés + sens dans commentaire
- Acronymes : transcrits si non dans dico
- Néologisme, verlan, argot : transcrits quand possible … sinon commentaire
- Format des heures : harmonisation
- Formes abrégées ambiguës : quand le contexte ne permet pas de lever => tel quel + commentaire
- Lettres répétées : 3 maintenues
- Casse : maintenu tel quel sauf si sms entier en maj
- Accent, cédille : rétablis
- Nombres : gardés si quantité, remplacés si utilisé pour son
- .....

Alpes4science : Interface transcription



Alpes4science : Quelques perles

Y doivent se fendre la poire en lisant les sms!:)

pense bien à boire … de l'eau

joie yeux ane y verre cerf col egg! :-P Gros bisous ++

vengeance sera faite espèce de lapin crétin fan de Dora !!!

Merci !