

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

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

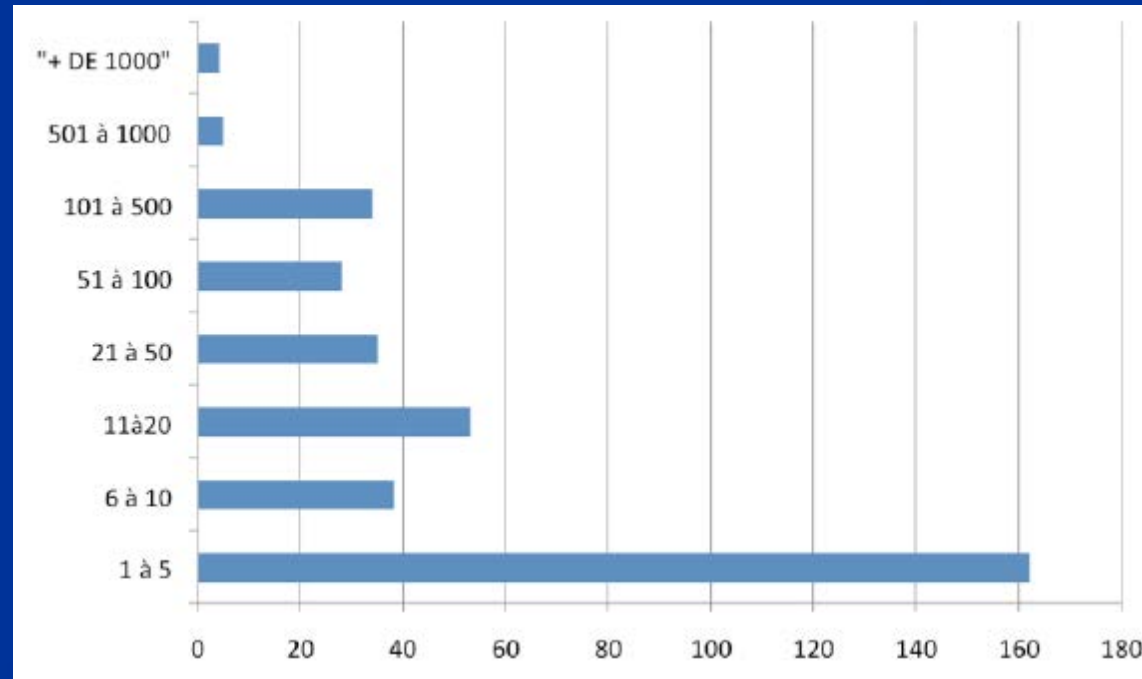
- Les opérateurs français ne sont pas intéressés par le projet
- Les numéros vert pour les portables n'existent pas en France, coût pour les donateurs
- Il faut payer aussi pour chaque sms reçu !
- Solutions payantes possibles via entreprises spécialisés ou opérateurs téléphoniques
- Plateforme ORANGE grâce au CG05

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

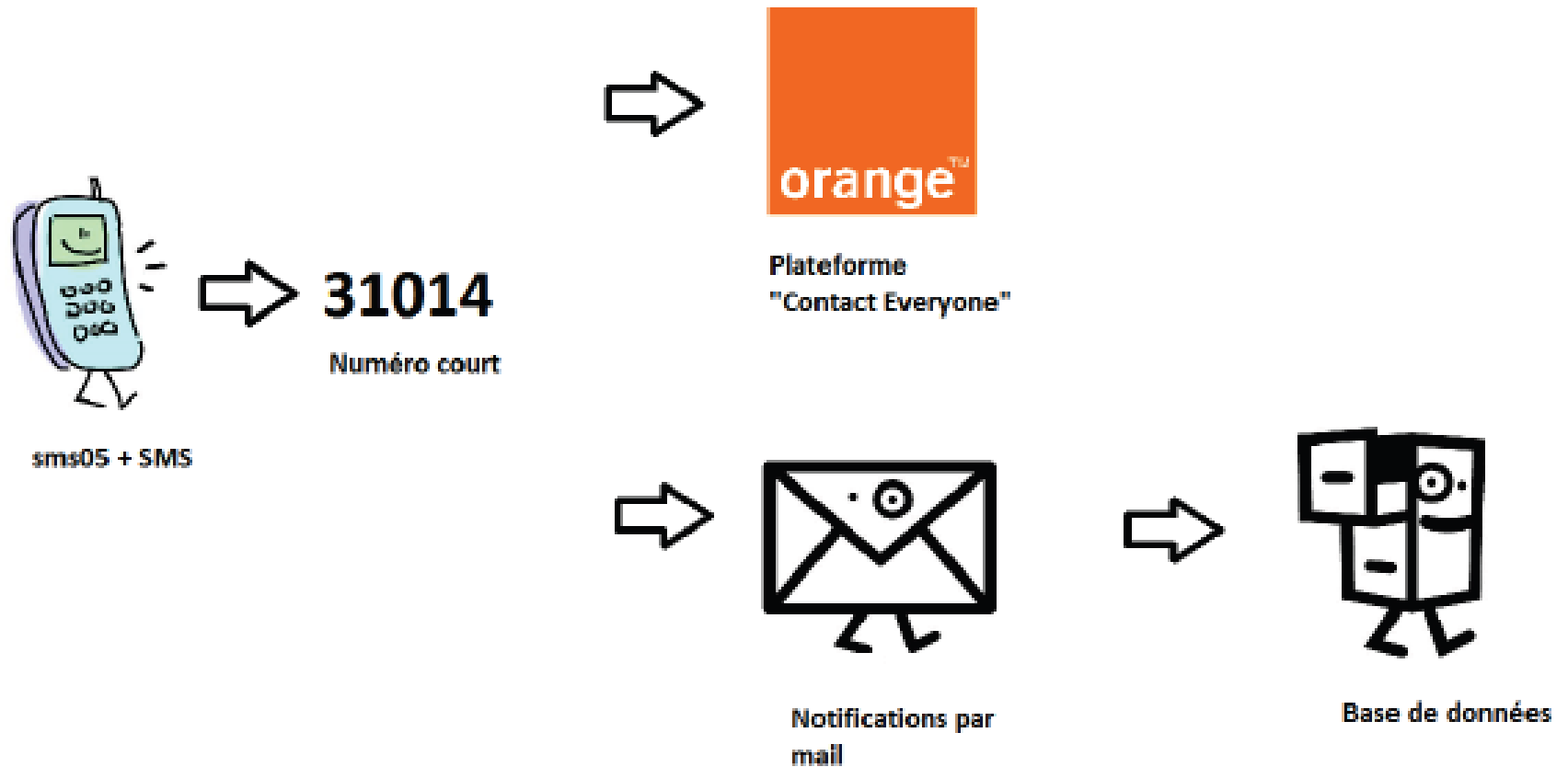
- Campagne de communication, politique de cadeaux (~0,25€/sms reçu)

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



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 penser, plutôt **SG** en **Guenièvre** ;-)

Alpes4science : Interface d'anonymisation

Texte du SMS 4825: Ok nō prob !!! I'll be There in 15 !!! Cheers bro !!!

[Déjà anonyme!](#)

[Choisir!](#)

Texte du SMS 4826: You filthy guy !!! Have a good night bro !!! I'm still working on That book !!! Don't eat thé savoury !!! It's not that good !!!

[Déjà anonyme!](#)

[Choisir!](#)

Texte du SMS 4828: yo ! Pié alias rodrigue est maintenant alias rené la taupe !!! Bizouxxx

[Déjà anonyme!](#)

[Choisir!](#)

SMS à anonymiser: hey ! Juste pour info tu peux appeler pié rené la taupe ! Love

rene ☐ Masquer ☐ Garder

Vous avez vu d'autres informations à anonymiser? Cliquez [ici](#)

Veillez les encadrer avec le symbole #

[Soumettre ces éléments](#)

Vous avez indiqué que 1 données supplémentaires sont à anonymiser

Explication de la donnée sélectionnée:

Choix du type de données pour "pié":

[Envoyer](#)

SMS avant modifications: hey ! Juste pour info tu peux appeler pié rené la taupe ! Love

ID: 4832

SMS avec modifications: hey ! Juste pour info tu peux appeler ***SURNOM_4*** rené la taupe ! Love

[Fermer](#)

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

Vous pouvez voir toutes les consignes de transcription [ici](#).

SMS à traduire: "coucou volci mon nouveau numero: ***TEL***"

Voici une proposition de découpage:

coucou volci mon nouveau numero ***TEL***

proposition de découpage du SMS

Valider ce découpage

Modifier ce découpage

outils de transcription

Pour garder le SMS tel quel, cocher cette case: ☐

coucou
volci
mon
nouveau
numero
TEL

zones de texte pour les transcriptions de chaque mot

☐ Garder ce mot tel quel

☐ Garder ce mot tel quel

☐ Garder ce mot tel quel

☐ Garder ce mot tel quel

☐ Garder ce mot tel quel

☐ Garder ce mot tel quel

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

... Traduction(s) possible(s) ...

numero

☐ Mot Inconnu

☐ Mot Inconnu

☐ Mot Inconnu

☐ Mot Inconnu

☐ Mot Inconnu

☐ Mot Inconnu

☐ Mot Inconnu

on pour ce terme

Veuillez laisser vos commentaires sur cette transcription ici:

zone de texte destinée aux commentaires

Envoyer

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 !

