Pour un ancrage contextuel de la statistique textuelle

Valérie Beaudouin
Télécom ParisTech

21 août 2009, Journée Alceste, Carcassonne

Contexte

- Retour d'expérience
  - une vingtaine d'années d'usage de la statistique textuelle et en particulier d'Alceste
- Bilan sur des usages
  - Dans des cadres théoriques et disciplinaires variés
  - Sur des corpus très différents
  - Dans des contextes institutionnels divers

Une question

- Pour quels types de corpus, la statistique textuelle peut-elle être une aide à la lecture ?
- Une question qui porte
  - sur les principes de constitution des corpus textuels et
  - sur la place centrale des variables du co(n)texte

Déroulement

- Quelques hypothèses
- Retour sur les corpus analysés à la lumière de cette hypothèse
  - Questions ouvertes
  - Collections de textes
  - Corpus de l'internet

Statistique textuelle adaptée

1. Pour des textes redondants, répétitifs -> illisibles en lecture linéaire
2. Pour des corpus de grande taille -> peu maîtrisables en lecture
3. Pour des textes avec des structures formelles difficiles à identifier en lecture linéaire
4. Pour des textes cohérents par leur contrat de communication

- assemblages de textes qui jouent sur la variation autour de l'identité (pôle du même)

Contrat de communication

- Qui parle ?
- Pour qui ?
- De quoi ?
- Pour quoi ?
- Comment ?
- Quand ?

- Émetteur
- Récepteur
- Thème
- Objectif
- Genre, forme, style
- Datation, chronologie

Contrôler les variations autour de ces critères

Les variables de contexte

- Les outils de statistique textuelle sont conçus pour intégrer les caractéristiques de la situation de communication :
  - Variables étoilées, variables illustratives…
- Les outils permettent :
  - d'identifier le vocabulaire spécifique de certaines situation spécifiques de communication;
  - De caractériser des regroupements de textes par ces situations.
- Ces variables expriment les lieux de variation dans la situation de communication

1. Réponses à des questions ouvertes
2. Collections homogènes
   - lettres de vœux
   - Portraits
   - Entretiens
   - Textes littéraires
3. Corpus de l'Internet :
   - Interactions (forums, chats, conversations avec machines…)
   - Sites Web

1. Réponses à des questions ouvertes

- Une situation de communication unique
- Une même question, un même destinataire, un même objectif, une même forme
- Un seul élément de variation : le répondant, l'auteur
- -> l'analyse va chercher à mettre en relation les types de réponses avec les caractéristiques des répondants

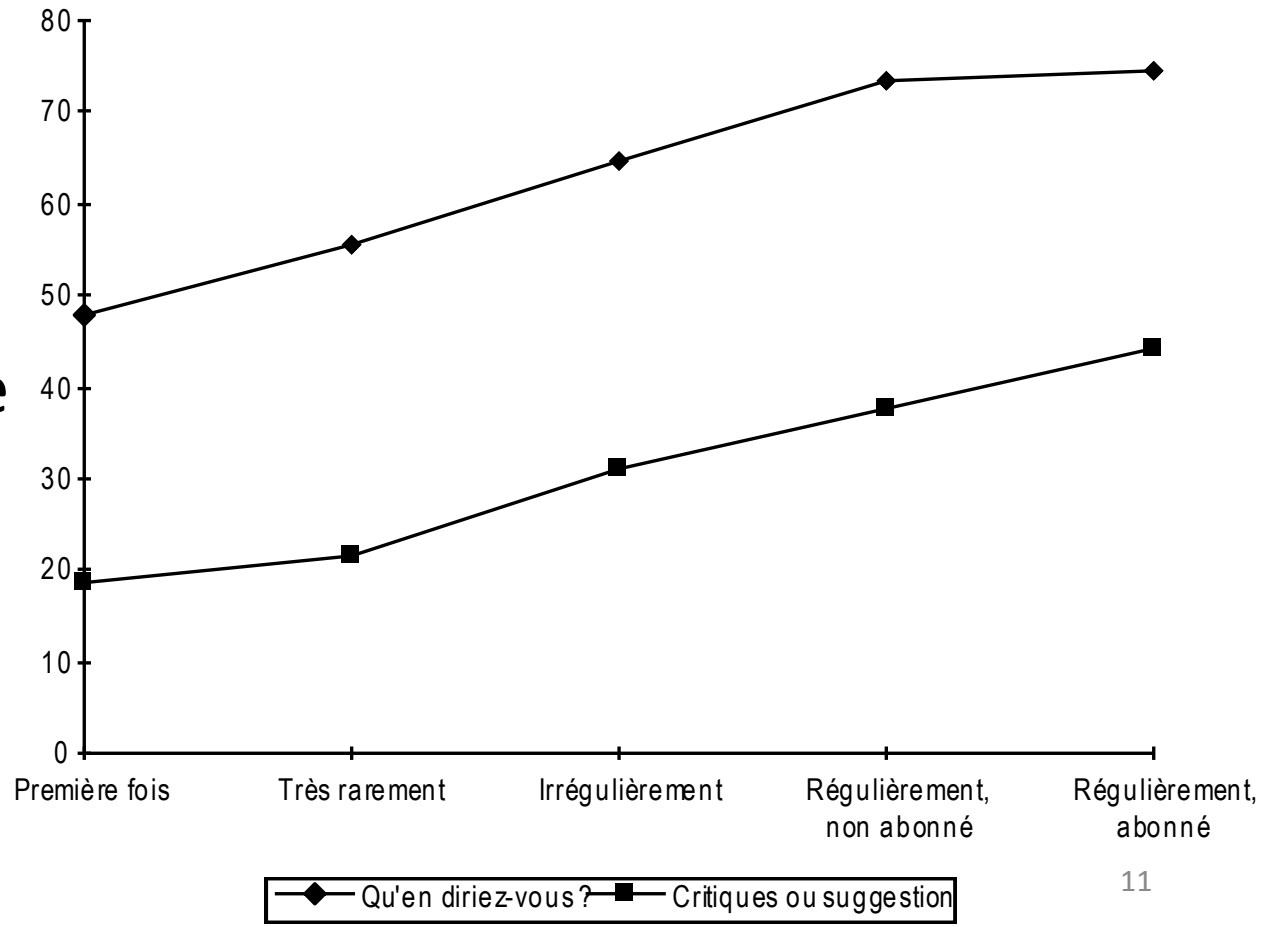
Tenir compte des conditions de production des réponses (1)

- La façon de répondre change selon le mode de questionnement

  – Si je vous dis petit déjeuner idéal, à quoi pensez-vous ?
    - Autoadministré : 38% « produits », 62% « contexte »
    - Téléphone : 95% « produits », 5% « contexte »

Tenir compte des conditions de production des réponses (2)

- Le taux de réponse varie selon la proximité avec la pratique
- Les publics de la Comédie Française



Tenir compte des conditions de production des réponses (3)

- La manière de répondre varie selon l'âge et le niveau de formation
- Si je vous dis x de qualité, quels sont les premiers mots qui vous viennent à l'esprit ?
  - Réponses sous forme de critères (en intension)
  - Réponses sous forme de prototypes (en extension)

2. Les collections de textes

- Quatre types de corpus
  - Un siècle de vœux adressés par les Caisses d'épargne
    - Etude réalisée à la demande du CENCEP coordonnée par l'ANVIE
    - Une rubrique annuelle du Journal des Caisses d'épargne paru de 1907 à 1982.
    - Objectif : repérer les valeurs des Caisses d'épargne et leur évolution au cours du temps. « Les entreprises ont-elles une âme ? »
  - 580 parcours d'insertion de jeunes en difficultés
    - Rédigés par les conseillers des Missions locales et Paio
    - Fêter les 10 ans des Missions locales
    - Un canevas unique
    - Parcours non représentatifs
    - Objectif : construire une typologie des parcours
  - 24 entretiens sur l'attitude face au risque du sida
    - Comprendre les freins aux changements de comportement face au risque
  - 46 pièces de théâtre de Corneille et Racine
    - Comprendre le lien entre structure lexico-sémantique et structure rythmique

Des collections de textes relevant de contrats proches

	Qui ?	Pour qui ?	Pour quoi ?	Comment ?	Quand ?
Journal des Caisses d'épargne	La Direction des Caisses d'épargne	Employés des Caisses d'épargne	Message sur la vision stratégique	Lettre annuelle	Un siècle, un texte par an
Portraits de jeunes	Des conseillers qui décrivent le parcours de jeunes	Ministère, Missions locales	Montrer l'utilité de l'institution	Cadre formel contraint	Une coupe en 1992
Entretiens	Echantillon diversifié	Sociologues	Répondre aux questions de l'enquêteur	Interview avec grille d'entretien	Une coupe
Théâtre de Corneille et Racine	Deux auteurs	Public de théâtre		Genre très codé, sous-genres : comédie/tragédie	Plus de 50 ans

CAISSE D'ÉPARGNE
DE 1907 À 1949

1. Lexique institutionnel : consolider le réseau (1906-1913)

conférence, commission supérieure, conférence générale, réuni, président, commission, membre, projet, session, discussion, éminent, bureau, question, groupement, divers, loi, émis, révision, voeu, région, autorité, parlement.

2. Lexique défensif (la polémique sur les excédents) : Caisse d'épargne ou banque de dépôt : perte de l'identité ? (1925-1930)

dépôt, maximum, commerçant, explication, revalorisation, augmentation, montant, petit, courir., simple, intérêt<, contester, déposant, guichets, changement, effectuer, afflux, capital, relevé, attirer, prétendre., statistique,

3. Lexique "gestionnaire" : périodes de trouble, les résultats priment sur la mission (après crise, après guerre)

excédent, versement, retrait, résultat, chiffre, opération, élévation, constatation, période, prospérité, atteint, quinzaine, solde, cours, crise.

4. Lexique affectif (défense des valeurs) : l'accomplissement de l'oeuvre commune ou le dévouement à la cause de l'épargne

dévoue+ (dévoué, dévouement), numéro, organisation, personnel, oeuvre, riche, renseignement, français, service, France, rôle, adresse, utilité.

Caisses d'épargne
De 1950 à nos jours

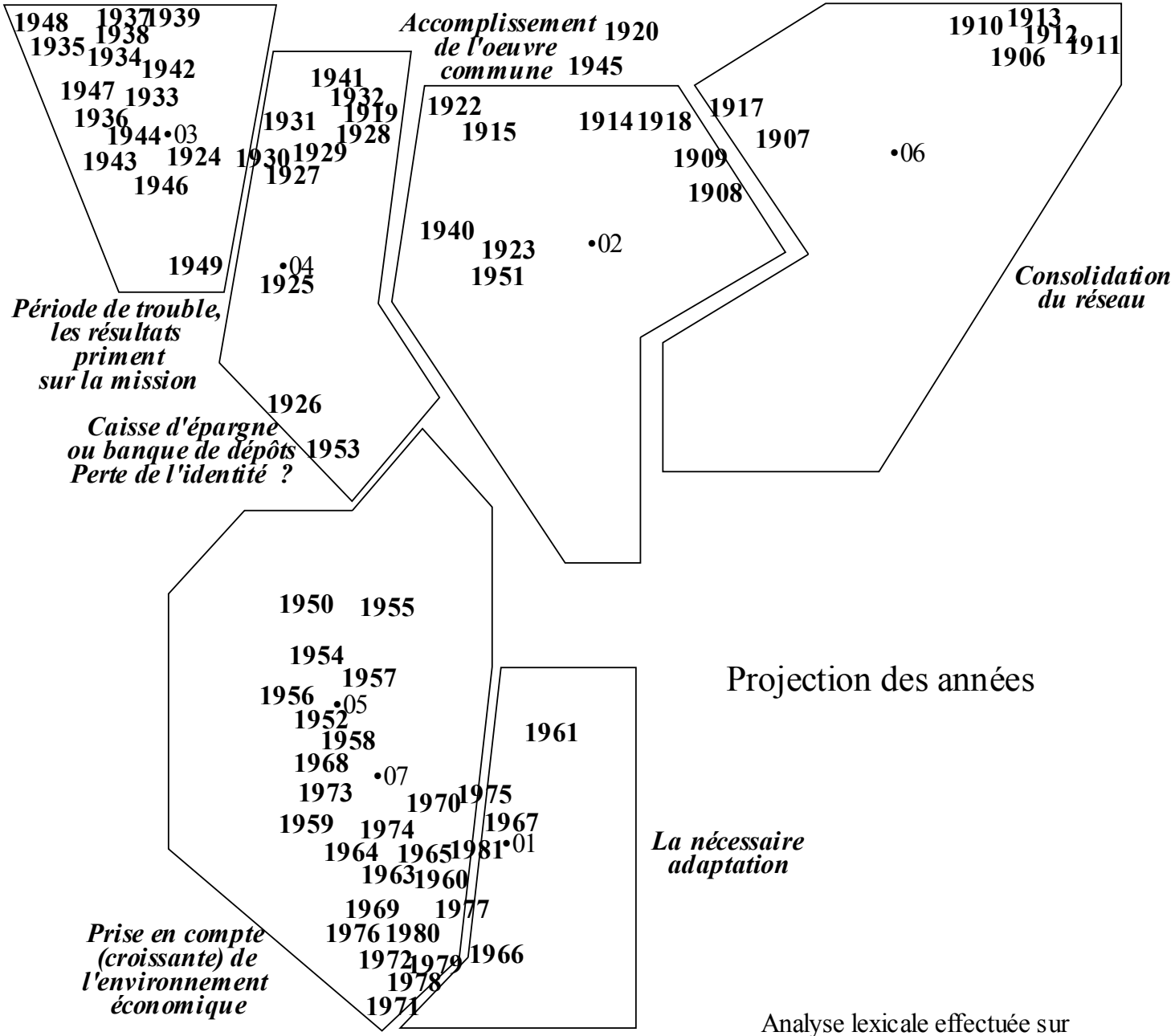
1. Lexique économique : prise en compte croissante de l'environnement économique

inflation, production, stabilité, prix, expansion, économie, international, plan<, devises, facteur, conjoncture, monétaire, hausse, psychologique, industrie, tension.

ménage, revenu, brut, disponible, taux, finance, placement, hypothèse, logement, consommation, liquide, forme, capacité, volume, fixation, léger, plafond, estimation, public, collecte, compte, indice, titre, préférence.

2. Lexique lié au changement : la nécessaire adaptation

problème, congrès national, adaptation, impérative, préoccupation, épargne logement, solution, mettre., complément, réseau, structure, utilisation, réflexion, prêt, agir., aller., service.



Analyse lexicale effectuée sur
"la revue de l'année"
(rubrique du Journal des Caisses d'épargne)
de 1907 à 1982

Exemple de portrait

Mission Locale des Trois Vallées (91)
Laurent
24 ans

Laurent vit chez ses parents mais connaît de grandes difficultés relationnelles avec son frère. Il a abandonné sa scolarité en classe de 3e et a fait jusque-là, des petits boulots et travaux précaires. Son projet professionnel est orienté vers les métiers du spectacle après avoir travaillé dans divers petits lieux musicaux et théâtraux.

Mars 1990 - juin 1991.

Laurent est motivé et déterminé lorsqu'il se rend pour la première fois à la Mission Locale. il entre en formation qualifiante dans le spectacle et sort de formation en juin 91.

La Mission repère chez Laurent une situation de fragilité sociale qui semble s'aggraver en raison de l'absence de débouchés professionnels correspondant à la formation acquise. Mais Laurent persiste dans son projet et la mission lui propose de poursuivre ses recherches d'emploi.

Avril 1992

Laurent disparaît et ne reprend contact qu'en avril 1992. Ses recherches ont été vaines et sa situation personnelle s'est dégradée : rupture avec le milieu familial, sans domicile fixe, problèmes de toxicomanie (héroïne) et alcoolisme.

Mai 1992

Après de nombreux entretiens à la Mission Locale Laurent rencontre une association spécialisée dans la toxicomanie des jeunes.

Il accepte de participer à un sevrage en juillet et une expérience voile en août (organisée par le Club de prévention et la Mission Locale).

Septembre 1992

Il lui est proposé un emploi de barman qui lui redonne confiance .

Octobre 1992

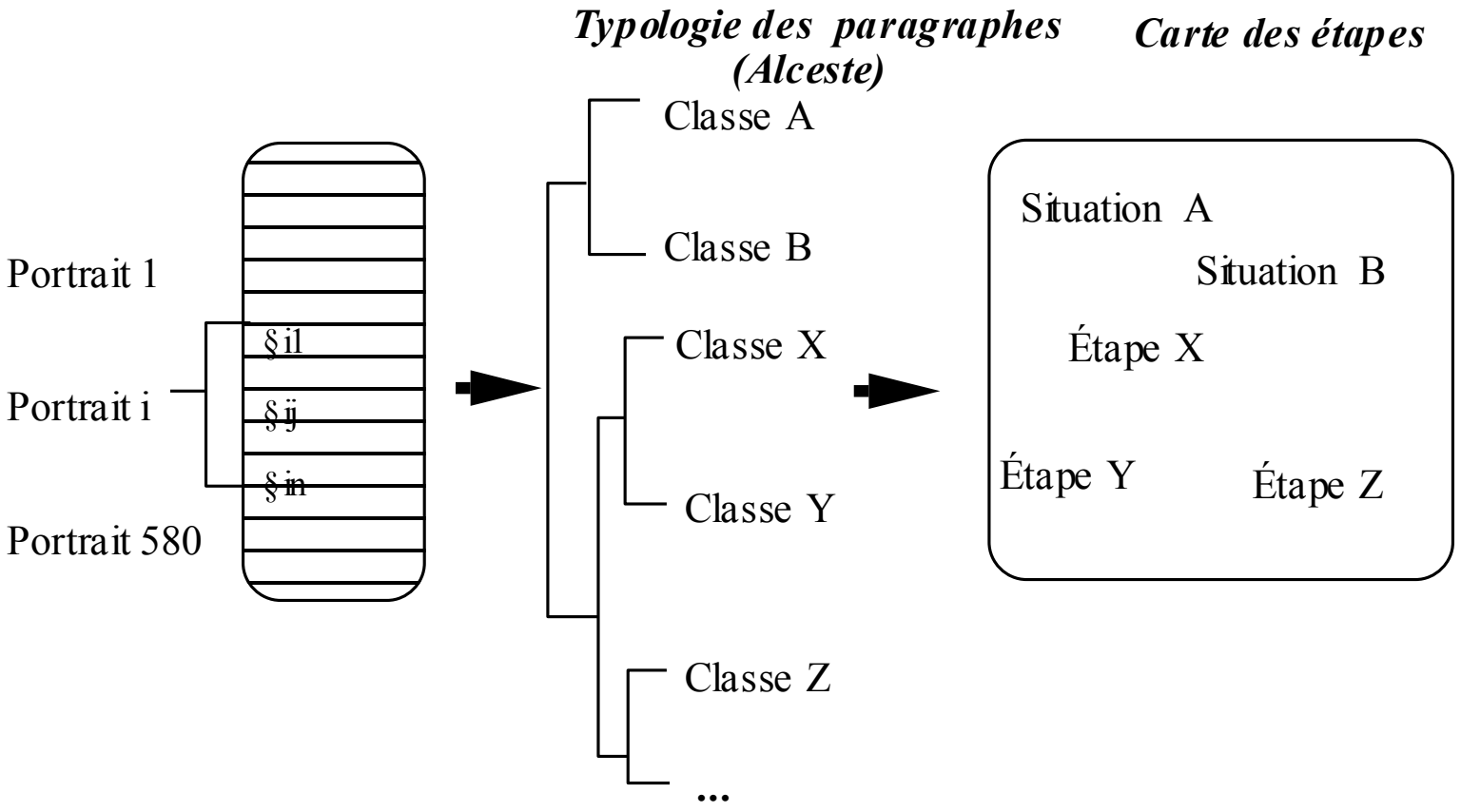
il passe les tests d'entrée aux Docks de France sur présentation de la Mission Locale.

•••

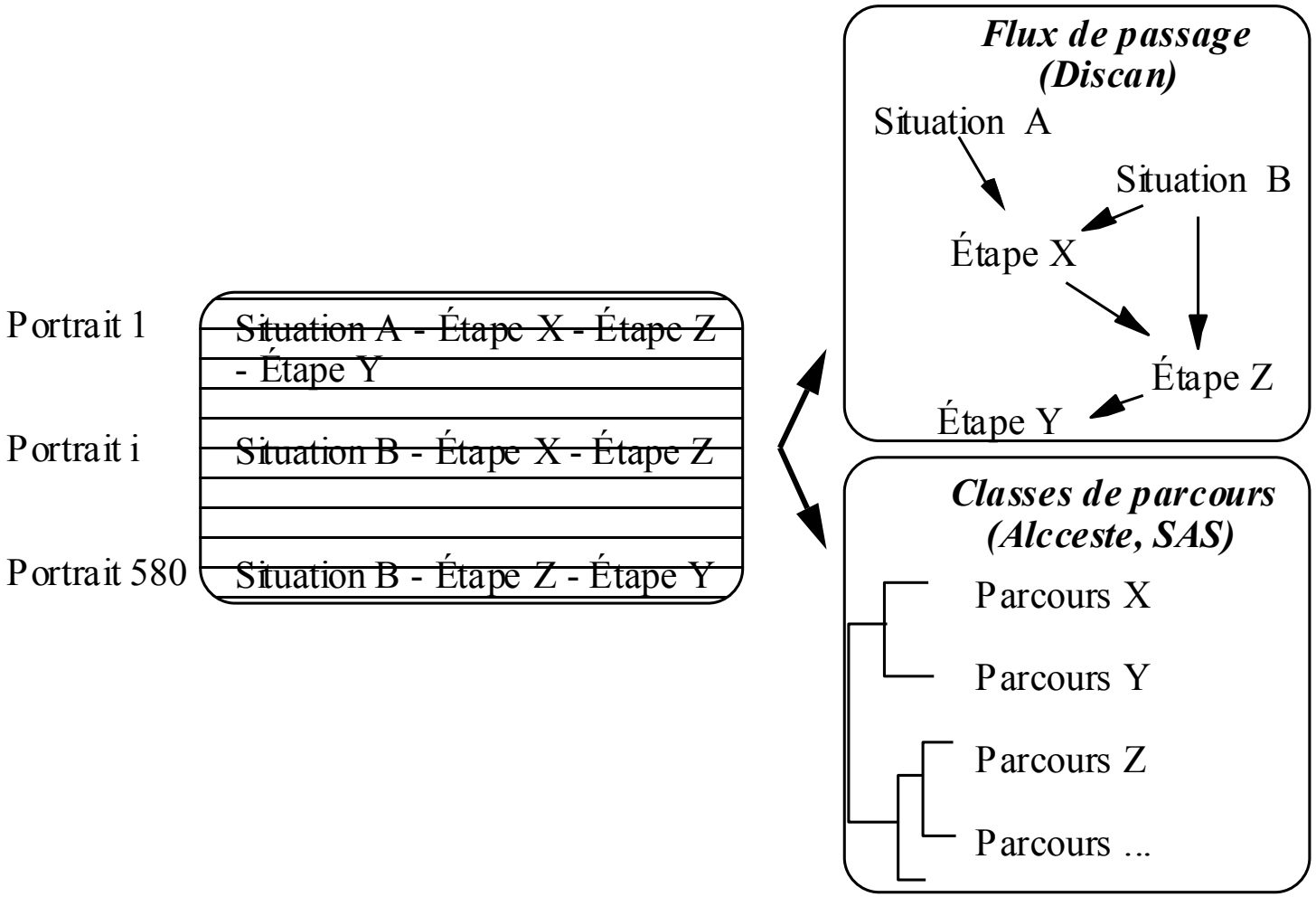
A ce jour Laurent est embauché sous contrat à durée indéterminée par la même entreprise.

Il revit chez ses parents. Il reste suivi par la Mission Locale. La relation de confiance établie entre la Mission Locale, Laurent, sa famille, et une qualité de partenariat local adapté ont permis à ce jeune de s'insérer professionnelle-ment.

1) REPÉRER LES ÉTAPES DES PARCOURS

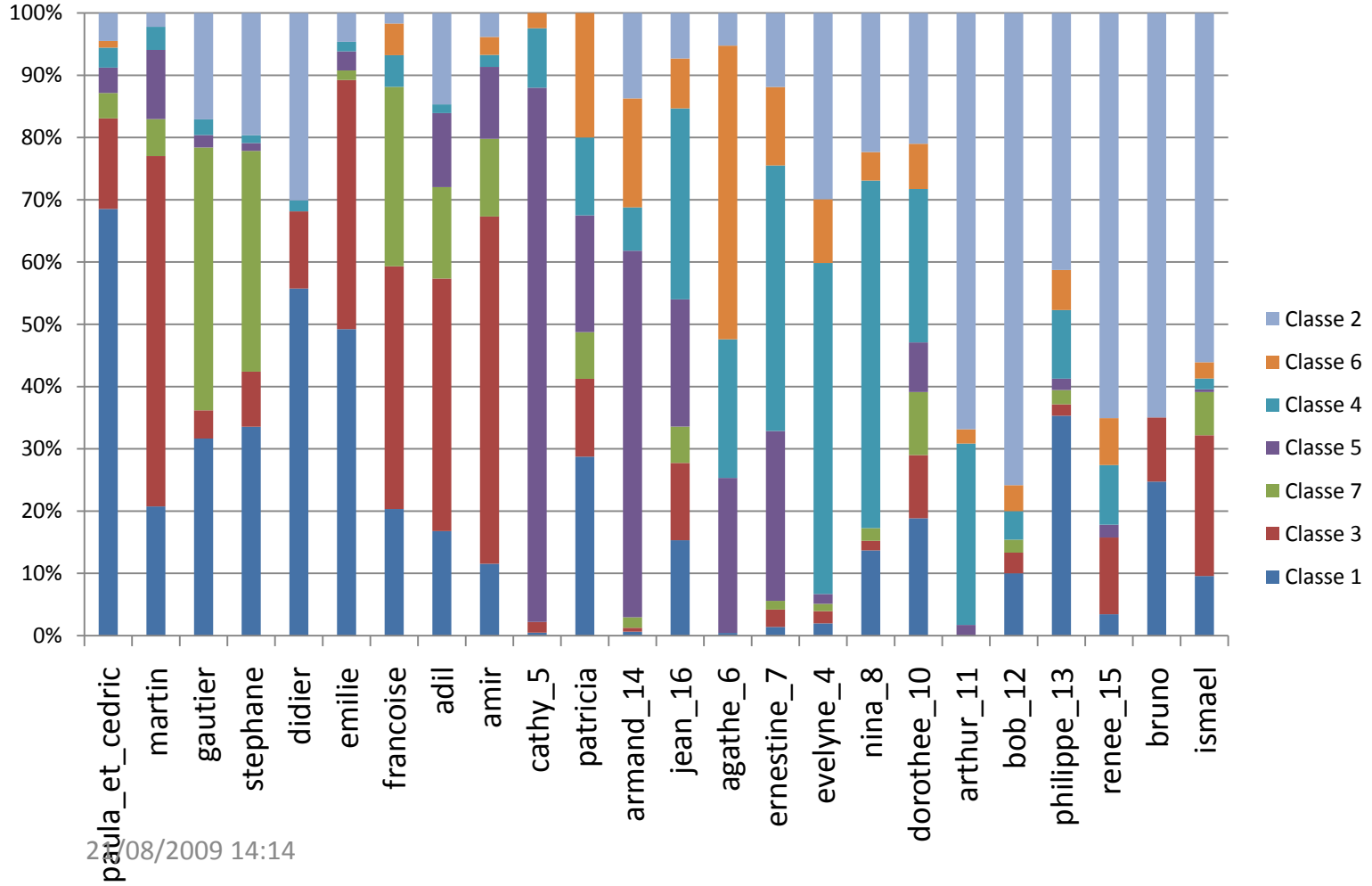


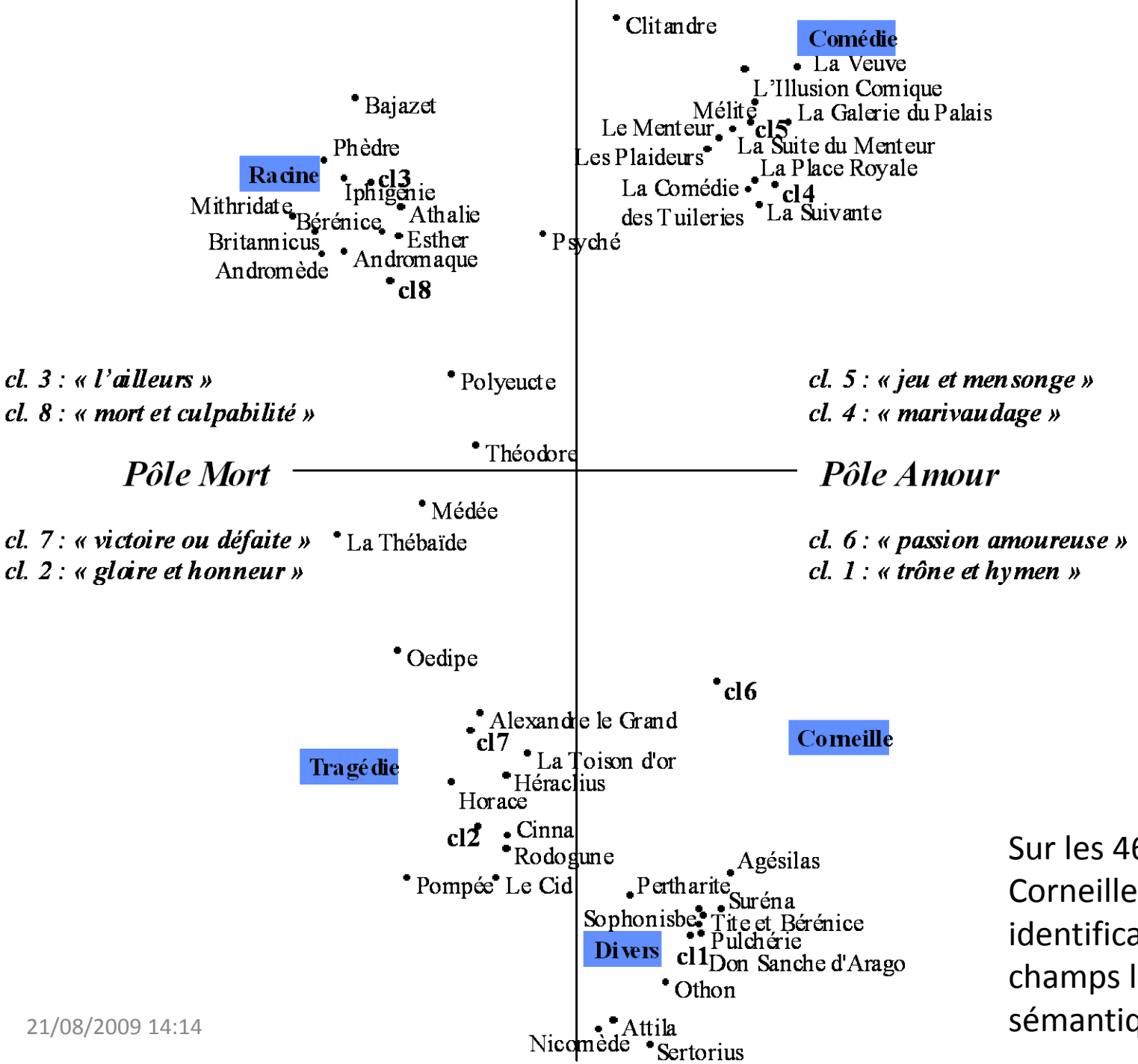
2) Reconstruire les parcours



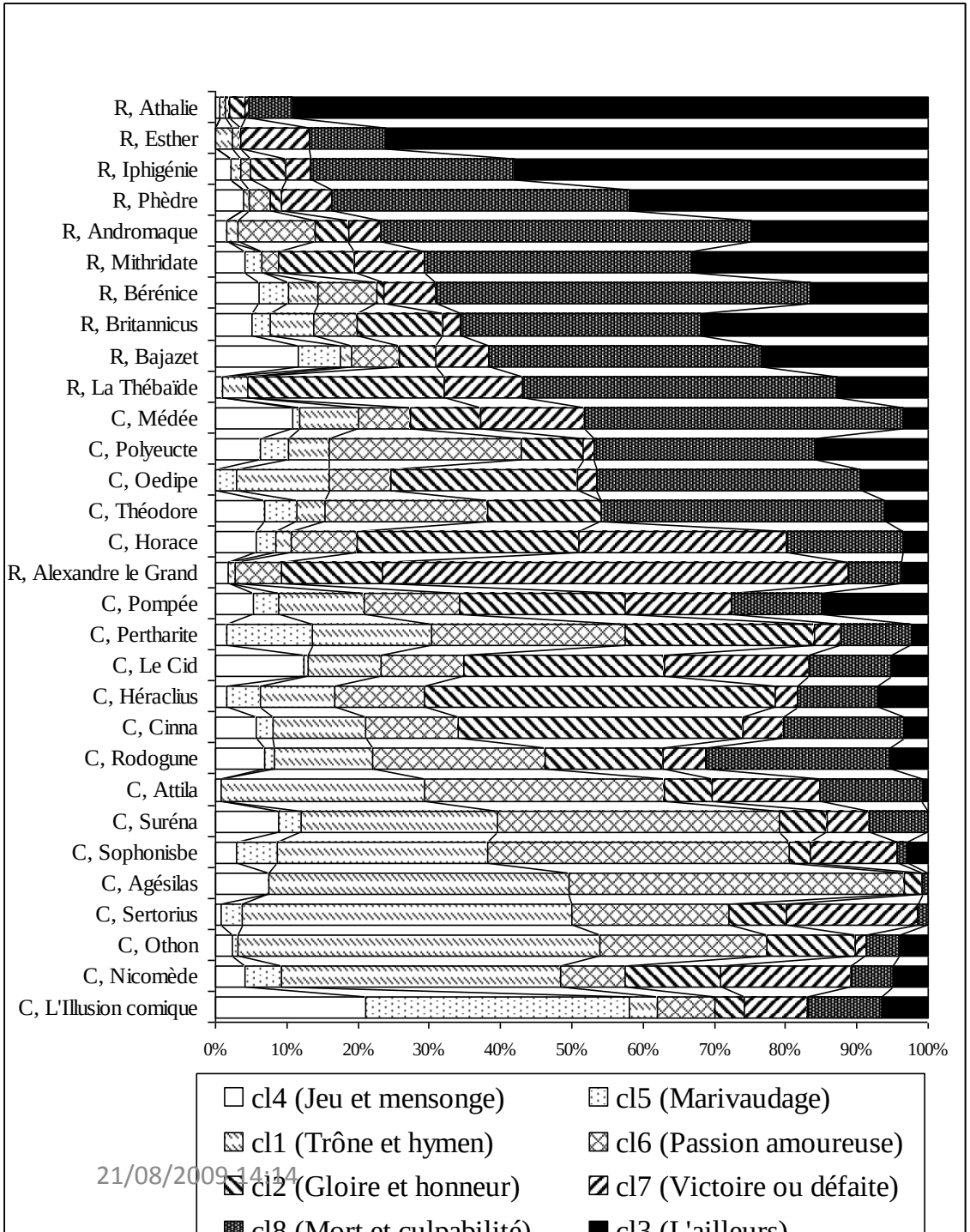
Attitude face aux risques du SIDA

Classe 1	peur, angoisse, résultat test	19%
Classe 2	relation, sexe, amour, amitié, fidélité/tromperie	23%
Classe 3	discussion, relation, préservatif, copain	11%
Classe 4	projet, vivre ensemble, quitter	14%
Classe 5	rentrer, sortir	19%
Classe 6	divorce, séparation, enfant, juge	7%
Classe 7	homosexualité, droqgue, milieu, précaution	6%





Sur les 46 pièces de Corneille et Racine, identification des champs lexico-sémantiques



- Tragédies de Corneille et Racine

Bilan sur les collections de texte

- Si la structure formelle est visible à l'œil nu, l'analyse fera émerger cette structure
  - Étapes de parcours pour les portraits
  - Thèmes de la grille d'entretien pour les entretiens
- -> nécessité d'aller plus loin
  - Explorer la place occupée par chaque thème
  - Travailler sur des sous-corpus
  - Déployer d'autres méthodes

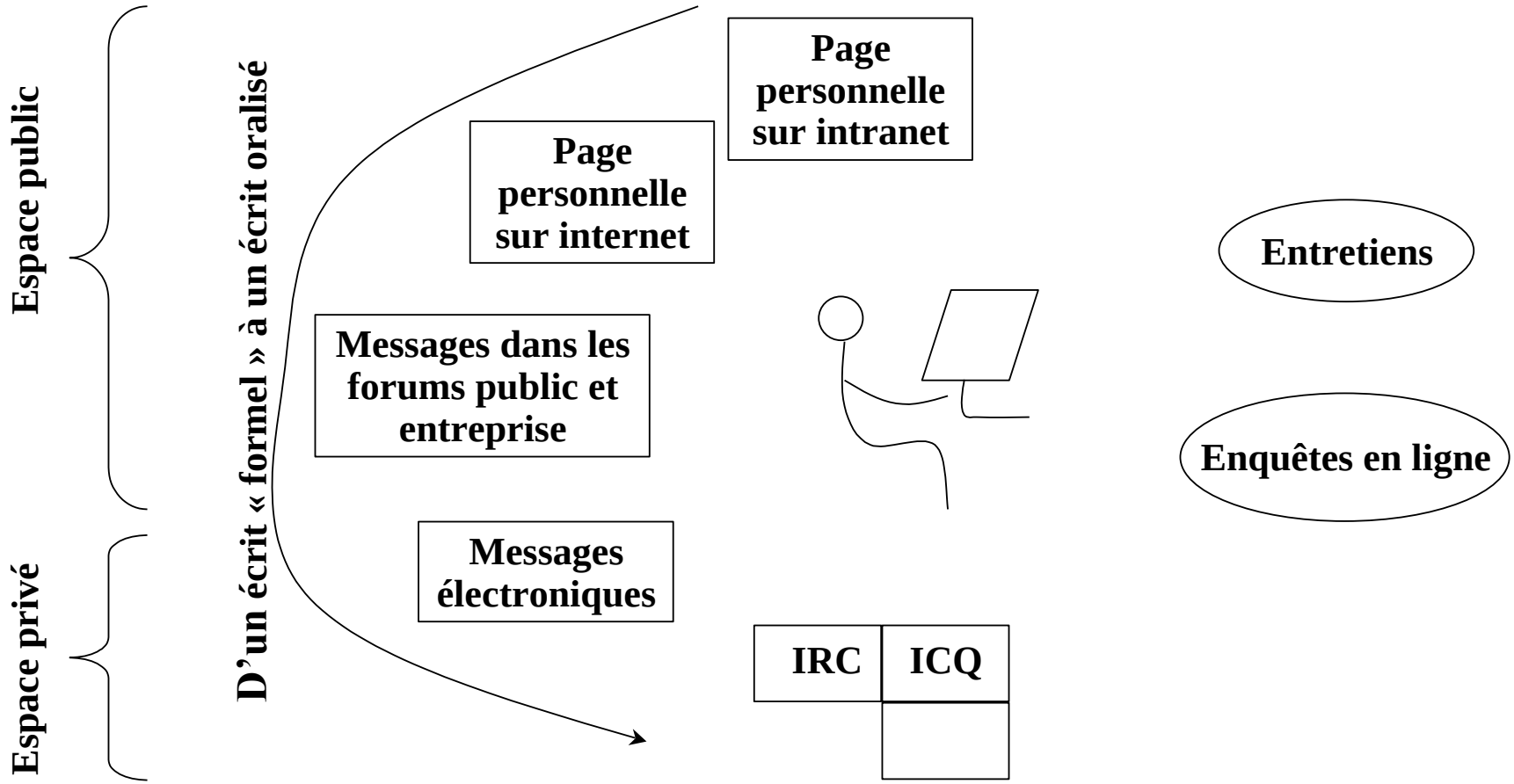
3. Les corpus liés aux services internet

- Le réseau comme ressource pour la constitution de corpus de textes (réservoir de textes numérisés)

- Corpus d'interactions (forums, chat…)
- Corpus d'espaces d'autopublication (sites personnels, blogs…)
- Et les espaces hybrides ?

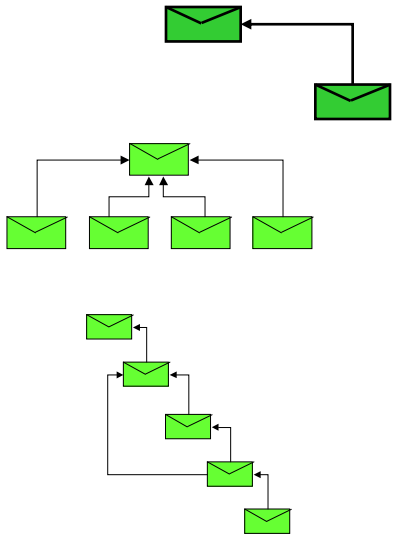
- Trois exemples : forums, chat, pages personnelles

Internet : promesses et désillusions

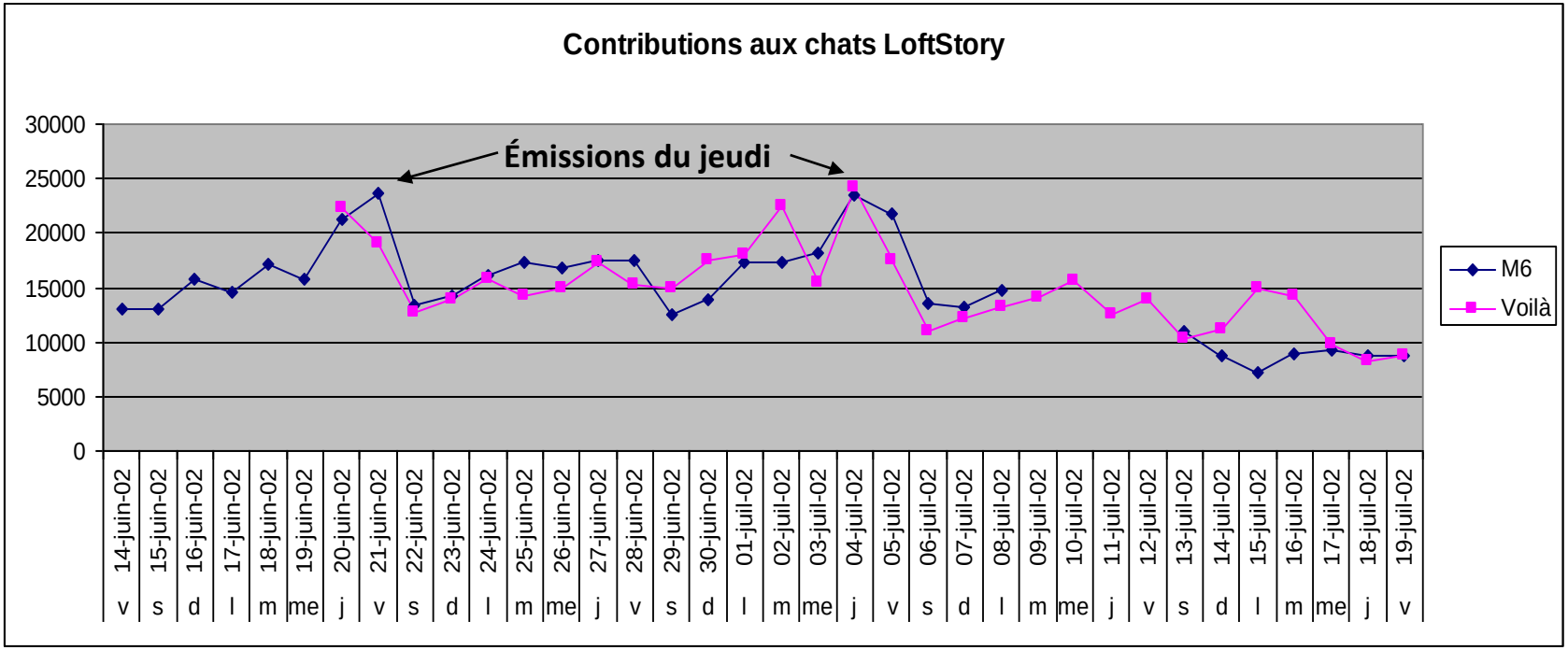


Les types d’activité dans un forum

- Construction d ’une typologie des messages qui renvoie à des types d’activité caractérisés
  - par un thème
  - par une structure d’interaction particulière
- Trois types d ’activité
  - échange technique (question-réponses)
    - 75% des messages
  - rappel des règles de conduite :
    - 10% des messages
  - surenchères humoristiques
    - 15% des messages

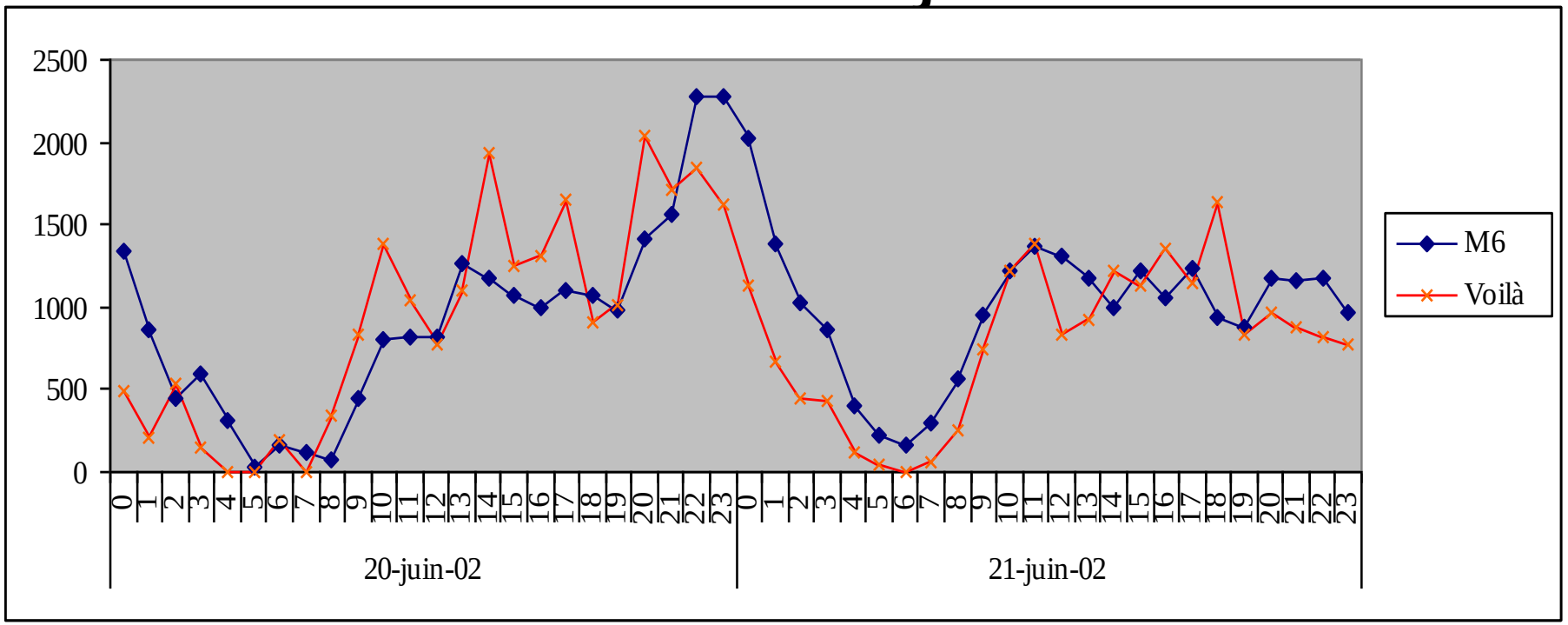


Le rythme des chats suit les grands moments de l'émission



- Le nombre de contributions sur les chats M6 et Voilà est globalement identique
- L'activité sur les chats M6 et Voilà suit le rythme de l'émission :
  - Forte augmentation de l'activité autour de l'émission du jeudi soir
  - Déclin progressif de l'activité après la clôture de l'émission

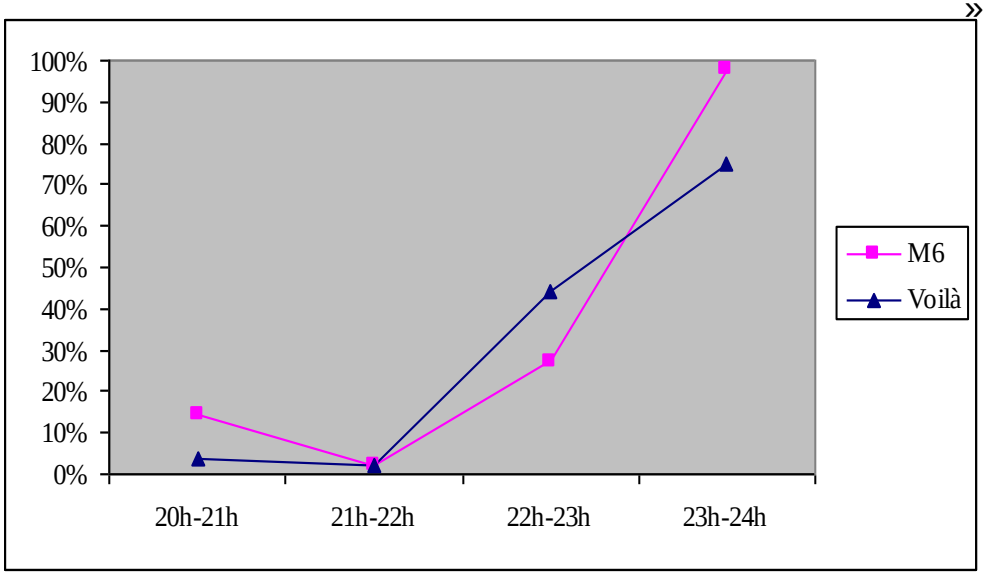
Emission du 20 juin 2002



- Sortie de Kamel, alors que David était le plus bas dans les votes tout au long de l'émission
- Pic d'activité au moment de l'annonce de la sortie de Kamel, mobilisation plus intense sur M6 que sur Voila

Part des messages critiques

- Identification des messages critiques par rapport à M6



» Vocabulaire spécifique : huissier+(37), truqu+er(77), hont+e(69), felic+(215), m6(136), magouille+(54), mauvais+(43), trich+er(42), boire.(25), castald+(37), prod+(48), chaine+(22), femme+(27), degout+er(25), denonc+er(15), perdre.(29), arnaque+(20), boouh(15), boouuhh(15), boycott+(17), dody(14), perdant+(15), zw(14), decu+(13), expressi+f(8), prochain+(14), producti+f(8), suisse+(10), vainqueur+(9), jaune+(12), france(23), chat+(35), colere+(10), debut+(14), facon+(12), gens(36), liberte+(8), merde+(39), plateau+(13), preuve+(9), truc+(19), yeux(12), pa+yer(16), prevoir.(10), regard+er(53), fait(109), possi+ble(18), travail<(19), boycot+(13), castelli(8), encule+(8), filliale(9), hyppo(14), kick+(23), laur+(11), lofteuse91(9), mariesolange(10), mumu(16), pf(8), pognon+(11), polio+(9), sylvano(15), tele(18), tiny(10), comique+(7), droit+(16), fina+l(32), hilare+(7), integre+(5), italien+(5), parti+(19), plein+(15), premier+(12), vive+(81), audience(5), avion+(7), bebe+(7);

- Part des messages critiques pendant la soirée du 20 juin
- Sur le chat M6, ces messages occupent l'intégralité des échanges à de 22h45, sur Voila le phénomène est un peu moins intense.

Construire une typologie des pages personnelles

- 100 000 pages vues par un panel d'internautes (panel NetValue-Nielsen//Netratings)

- Pages décrites
  - par des traits hypertextuels (liens, images, traits méta)
  - par le vocabulaire (taille du vocabulaire et place des pronoms)

- Une chaine de traitement complexe pour passer de la page à sa description par des traits
  - projet RNRT SensNet : FT, NetRatings, LIMSI, PIII
  - Beaudouin V., Fleury S., Pasquier M., Habert B. & Licoppe C. (2002). "Décrire la toile pour mieux comprendre les parcours. Sites personnels et sites marchands", Réseaux, Vol. 20, n°116, p. p. 19-51.

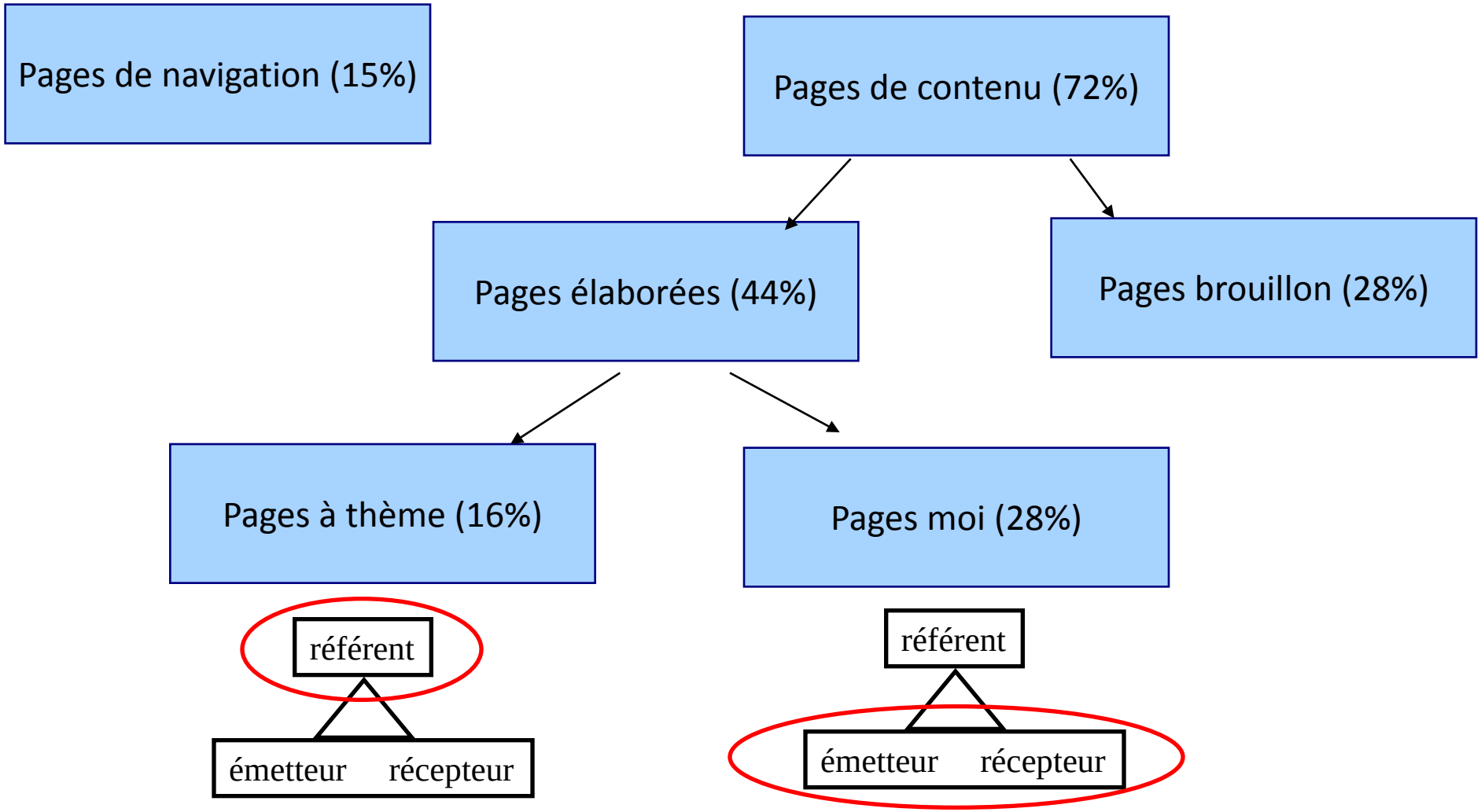
Les liens hypermedia selon le serveur d'hébergement

Serveur d'hébergement	Nb moyen de pages visitéees par site	Nb moyen de liens internes	Nb moyen de liens externes	Nb moyen d'images	Nb moyen d'images externes	Nb moyen de liens vers BAL
free_fr	6,4	20,1	3,5	7,0	0,8	0,4
perso.wanadoo.fr	5,4	9,6	1,9	5,1	0,3	0,5
perso.club-interne	5,3	13,8	3,6	7,1	1,4	0,7
www.chez.com	5,0	19,0	4,2	5,7	0,6	0,9
le-village.ifrance.c	4,8	7,0	3,1	5,0	0,6	0,8
www.multimania.c	4,8	6,1	3,7	4,6	0,5	0,4
ifrance.com	3,9	5,7	3,5	5,8	0,5	0,5
www.geocities.com	3,0	6,5	6,5	6,7	3,4	0,4
autres	3,2	10,3	4,5	6,5	0,9	0,5

Rq : sites perso wanadoo visitées 5,4 pages en moyenne, 1,9 liens externes, 9,6 liens internes

Une très faible part des sites est effectivement visitée

Typologie des pages personnelles



Pages visitées : types de pages

- Les types de pages sont caractérisés par des configurations spécifiques de traits.
- Les types peuvent être interprétés comme des états différents d'élaboration de la page
  - Pages sophistiquées: beaucoup de liens, beaucoup d'images
    - pages d'experts avec maîtrise de l'écriture html, sans pronoms
    - pages avec beaucoup de pronoms personnels, liens vers boîte aux lettres
  - Pages "brouillon" : peu de liens, peu d'images, peu de pronoms, peu de maîtrise dans la définition de la forme de la page
- On mesure plutôt le degré d'élaboration de la page que des sous-genres au sein des sites personnels.

Bilan sur les corpus liés aux services internet

- Les espaces d'interactions : adaptés à la stat textuelle
  - espaces normés (netiquette)
  - règles locales de fonctionnement
  - formats d'intervention contraints
  - Le texte domine
- Les espaces d'autopublication : problématiques
  - Premier genre numérique, mais
  - Genre instable, en transformation permanente
    - Contenu et Forme
  - Genre multimédia et non pas textuel
  - Variabilité sur l'émetteur, sur le public, sur l'objectif, sur le genre
- Et les sites de réseaux sociaux : nouvelle frontière pour la recherche

Conclusion

- Le nerf de la guerre tient à la constitution du corpus
- S'interroger sur le contrat de communication et limiter les formes de variation
- Enrichir la description du texte avec des variables illustratives, qui décrivent les caractéristiques du contrat de communication

Bibliographie indicative

- Beaudouin V. & Lahlou S. (1993). *L'analyse lexicale : outil d'exploration des représentations. Paris, CRÉDOC, Cahier de Recherche, 146 p.*
- Benzécri J.-P. é. (1981). Introduction I. *In, Pratique de l'analyse des données, Linguistique et lexicologie. Paris, Dunod.*
- Benzécri J.-P. é. (1981). *Pratique de l'analyse des données, Linguistique et lexicologie. Paris, Dunodp.*
- Brunet É. (1993). *Un hypertexte statistique : Hyperbase. JADT 1993 : Actes des Secondes journées internationales d'analyse statistique de données textuelles, Montpellier, ENST-Telecom, xxx.*
- Habert B., Nazarenko A. & Salem A. (1997). *Les linguistiques de corpus. Paris, Armand Colin/Masson, 240 p. p.*
- Lahlou S. (1995). *Vers une théorie de l'interprétation en analyse statistique des données textuelles. JADT 1995 : III Giornate internazionali di Analisi Statistica dei Dati Testuali, Roma, CISU, 221-228.*
- Lebart L. & Salem A. (1988). *Analyse statistique des données textuelles. Paris, Dunod, xxx p.*
- Lebart L. & Salem A. (1994). *Statistique textuelle. Paris, Dunod, 342 p. p.*
- Muller C. (1967, 1992). *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille. Paris, Larousse, 1967, réimpression aux éditions Slatkine, 1979, 1992 382 p.*
- Muller C. (1977, 1992). *Principes et méthodes de statistique lexicale, Larousse, 1977, réimpression Champion-Slatkine, 1992, 211 p.*
- Reinert M. (1990). "ALCESTE : Une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval", *Bulletin de méthodologie sociologique, n°26, p. 24-54.*
- Reinert M. (1993). "Les "mondes lexicaux" et leur logique", *Langage et société, n°66, p. 5-39.*
- Salem A. (xxx). "La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert", *4ème Colloque de lexicologie politique Langages de la révolution 1770-1815.*
- Zipf G. K. (1936, 1974). *La psychobiologie du langage : une introduction à la philologie dynamique. Paris, RETZ-CEPL, 232 p. p.*

Bibliographie : exemples présentés

- Beaudouin V., Hebel P., Lahlou S. & Le Bihan H. (1992). Enquête sur la qualité perçue. *In: L. A.-E. CRÉDOC, MIND MOVERS, Comprendre et évaluer la qualité. Paris, CRÉDOC, Cahier de Recherche. n °39.*
- Beaudouin V. & Lahlou S. (1993). *L'analyse lexicale : outil d'exploration des représentations. Paris, CRÉDOC, Cahier de Recherche, 146 p.*
- Beaudouin V., Lahlou S. & Yvon F. (1993). *Réponse à une question ouverte : incidence du mode de questionnement. JADT 1993 : Actes des Secondes journées internationales d'analyse statistique de données textuelles, Montpellier, ENST-Telecom, 131-142.*
- Beaudouin V. & Aucouturier A.-L. (1995). "Histoires d'insertion : analyse lexicale de 580 récits de parcours de jeunes", *Travail et emploi, n°65, p. 19-38.*
- Beaudouin V., Maresca B. & Guy J.-M. d. (1997). *Les publics de la Comédie-Française. Fréquentation et image de la salle Richelieu. Paris, La Documentation française, 288 p. p.*
- Beaudouin V. (1998). *Rythme et univers lexicaux chez Corneille et Racine. JADT 98 : 4èmes journées internationales d'analyse statistique des données textuelles, Nice, 85-93.*
- Beaudouin V. & Velkovska J. (1999). "Constitution d'un espace de communication sur Internet (Forums, pages personnelles, courrier électronique...)", *Internet, un nouveau mode de communication ?, 17, 97, p. 121-177.*
- Beaudouin V. (2002). *Mètre et rythmes du vers classique - Corneille et Racine. Paris, Champion, coll. Lettres numériques.p.*
- Beaudouin V., Beauvisage T., Cardon D. & Velkovska J. (2003). L'entrelacement des médias dans la constitution des publics de Loft Story, Issy-les-Moulineaux, France Télécom R&D, 64 p.
- Beaudouin V., Fleury S. & Pasquier M. (2004). Les pages personnelles comme terrain d'expérimentation. *In: F. Mourlhon-Dallies, F. Rakotonoelina and S. Reboul-Touré, Les carnets du Cediscor, Presses Sorbonne nouvelle. 8, 143-164.*