Conception et prototypage d'un outil web de médiation et d'aide au dialogue tchaté écrit en langue seconde

Achille Falaise

Université Joseph Fourier - Grenoble I

N° attribué par la bibliothèque

THÈSE

pour obtenir le grade de

Docteur de l'Université Joseph Fourier

Spécialité : **Informatique**

préparée au laboratoire **Laboratoire d'Informatique de Grenoble (LIG), Groupe d'Étude en Traduction Automatique et le Traitement Automatisé des Langues et de la Parole (GETALP)**

dans le cadre de l'École Doctorale **École Doctorale Mathématiques, Sciences et Technologie de l'Information, Informatique (EDMSTII)**

présentée et soutenue publiquement
par

Achille Falaise

le 25 septembre 2009

Titre:

Conception et prototypage d'un outil web de médiation et d'aide au dialogue tchaté écrit en langue seconde

Directeur de thèse: **Christian Boitet**
Codirecteur de thèse: **Hervé Blanchon**

Jury

M. Jean Caelen,	Président du jury
Mme. Violaine Prince,	Rapporteur
M. Patrice Pognan,	Rapporteur
M. Emmanuel Planas,	Examinateur
M. Rémi Marand,	Examinateur
M. Christian Boitet,	Examinateur
M. Hervé Blanchon,	Examinateur

Remerciements

Je souhaite remercier en premier lieu Violaine Prince et Patrice Pognan d'avoir accepté d'être les rapporteurs de cette thèse, malgré un emploi du temps déjà chargé. Je remercie Emmanuel Planas pour ses conseils, ainsi que Jean Caelen, qui ont bien voulu faire partie de mon jury.

Je remercie mes encadrants, Christian Boitet et Hervé Blanchon au GETALP, et Rémi Marand à Prosodie, pour avoir bien voulu m'accorder leur confiance et une grande liberté de travail, tout en ayant été répondu présent lorsque c'était nécessaire.

Je remercie Alain Bernard, sans qui cette thèse n'aurait pu exister.

Je remercie tous les chercheurs, thésards et stagiaires que j'ai pu rencontrer dans l'équipe GETALP, et plus largement au laboratoire LIG. Merci particulièrement à Georges Fafiotte pour ses discussions et ses conseils, à Valérie Bellynck, Laurent Besacier, Jean-Claude Durand, Jérôme Goulian, Jean-Philippe Guilbaud, Cong-Phap Huynh, Hong-Thai Nguyen, Marion Potet, David Rouquet, Jean-François Sérignat, Didier Schwab, Sopheap Seng, Gilles Sérasset, Dominique Vaufreydaz, pour leur aide, leurs conseils.

Merci aux membres du LIDILEM avec qui j'ai eu le plaisir de travailler et de discuter, en particulier Agnès Tutin, Francis Grossmann et Christelle Cavala.

Je remercie mes parents, Martine Lecomte et Gérard Falaise, pour m'avoir montré que la vie est avant tout une tentative de réaliser ses rêves. Merci à Bruno Jouniaux pour m'avoir montré que l'on pouvait aussi se réaliser par la science. Merci aux professeurs qui, tout au long de ma scolarité, m'ont donné les outils pour y parvenir.

Enfin, merci aux amis, aux joyeux thésards Cyrille Vézy, Fabrice Salpetrier, Stéphane Caro et Éléna Rauch, à tous les membres de la fabuleuse troupe Excalibur-Dauphiné, que je n'ai pas la place de tous lister ici ; une pensée spéciale toutefois, pour Annie Chevallier et son aide précieuse en cette fin de thèse.

Table des matières

La radio bracelet

«Le prisonnier a été emmené dans la Chambre d'Interrogatoire n°9.»
Il tira le verrou de la porte, entra dans la chambre. Il entendit la porte se refermer derrière lui.
«Fichez le camp !» dit le prisonnier, en souriant.
Le psychiatre fut frappé par son sourire. Il éclairait la chambre, c'était quelque chose d'ensoleillé, de chaud.

(...)

« Je viens vous aider », dit le psychiatre, en frissonnant.
Le prisonnier rit : « Vous voulez savoir la raison de ce calme ici ; je viens de mettre en pièces l'appareil de radio.»
« Un violent», se dit le docteur.
Le prisonnier lut dans sa pensée, sourit, tendit une main amicale. «Non, simplement pour faire taire ce bla-bla-bla.
- Vous êtes bien Albert Brock, qui se nomme lui-même le Criminel ?»
Brock approuva gentiment de la tête. «Avant de commencer...» Il avança d'un mouvement calme, et détacha rapidement la radio du médecin. Il la mordit, l'écrasant entre ses dents comme une noisette. «C'est mieux ainsi».
Le psychiatre fixa l'appareil détruit. «Vous accroissez dangereusement la note.
- Aucune importance, dit en souriant le patient.

(...)

- On commence ? demanda le psychiatre.
- Allons-y. La première victime (...), ce fut mon téléphone. Crime atroce. Je l'ai jeté dans le vide-ordures de la cuisine. (...) Puis, à coups de revolver, j'ai tué l'appareil de télévision !

(...)

- Et si vous me racontiez quand, pour la première fois, vous avez commencé à haïr le téléphone.

(...)

- Un de mes oncles l'appelait l'appareil fantôme. (...) Le téléphone m'a toujours semblé un instrument impersonnel. Si on se laisse aller, il absorbe votre personnalité à travers ses fils. Si vous résistez, il l'annihile et ce qu'on entend à l'autre bout, c'est une drôle de voix, faite d'acier, de cuivre, de matière plastique, sans chaleur, sans réalité. (...) Il est là et demande qu'on appelle quelqu'un qui n'a pas envie d'être appelé.

Les amis m'appelaient, m'appelaient, et m'appelaient encore. Un véritable enfer (...). Quand ce n'était pas le téléphone, c'était la télévision, la radio, le phonographe (...), les films au cinéma du coin de la rue, les réclames projetées jusque sur les nuages les plus bas. (...) Alors m'est venue l'idée de la machine diathermique portative. J'en ai loué une ; je l'ai prise avec moi dans l'autobus, au retour, ce soir là. Il y avait assis en rang tous ceux qui venaient de finir leur travail, morts de fatigue, avec leur petite radio à leur poignet, parlant à leur femme, leur disant «Je suis maintenant à la hauteur de la Quarante-troisième rue.» (...) Un mari jurait : «Sors de ce bar, nom de Dieu, et rentre préparer le dîner !». Alors...j'ai mis en marche mon appareil de diathermie ! Arrêt complet ! (...) Le silence ! Un terrible silence inattendu. Le usagers du bus avaient la possibilité de se parler les uns aux autres. Quelle panique ! Panique éperdue, animale !
- La police vous a arrêté ?
- L'autobus a dû s'arrêter. Après tout, (...) les maris et les femmes avaient perdu tout contact avec le réel. C'était le Pandémonium, l'émeute, le chaos. (...) Une équipe de secours arriva, m'encercla, me réprimanda, me mis à l'amende.

(...)

- Hmmm, fit le psychiatre.»
Il appuya sur un bouton de signalisation, la porte s'ouvrit, il sortit, la porte se referma d'elle-même, à nouveau verrouillée. Seul, il s'achemina à travers les bureaux, les couloirs. Il retourna à son bureau. Une sonnerie se fit entendre ; une voix descendit du plafonnier :
«Docteur ?
- Je viens justement de finir avec Brock, dit le psychiatre.
- Le diagnostic ?
- Il paraît complètement désorienté, mais sociable. Il refuse d'accepter les réalités courantes de la vie et de s'y conformer.»
Trois téléphones sonnèrent en même temps. Un radio-bracelet de rechange fit entendre son appel de sauterelle blessée dans un tiroir du bureau. (...) Le psychiatre, fredonnant calmement, mit le radio-bracelet à son poignet, parla un moment dans l'inter, leva le récepteur d'un des appareils, parla, prit un autre récepteur, parla, prit le troisième récepteur, parla ; il appuya sur le bouton de la radio-bracelet, parla ; il était calme, tranquille, le visage froid et serein. (...) Et il continua tranquillement de cette façon tout l'après-midi, dans la fraîcheur de l'air conditionné : téléphone, radio-bracelet, inter, téléphone, radio-bracelet, inter, téléphone, radio-bracelet, inter, téléphone, radio-bracelet, inter, téléphone, radio-bracelet, inter, téléphone, radio-bracelet...

Ray Bradbury, *The Murderer*, 1953
Traduction de l'américain par Richard Negrou, sous le titre *Le Criminel*, 1956

La traductrice

La montagne était déjà creusée d'une trentaine de galeries tout autour desquelles avaient été installés, au cœur vif de la glace, les entrepôts et les émetteurs radio et TV de l'Expédition Polaire Internationale, en abrégé l'E.P.I. C'était un beau nom. La ville dans la montagne se nommait EPI 1 et celle qui était abritée sous la glace du plateau 612 se nommait EPI 2. EPI 2 comprenait toutes les autres installations, et la pile atomique qui fournissait la force, la lumière et la chaleur aux deux villes protégées et à EPI 3 la ville de surface, composée des hangars, des véhicules et de toutes les machines qui attaquaient la glace de toutes les façons que la technique avait pu imaginer. Jamais une entreprise internationale d'une telle ampleur n'avait été réalisée. Il semblait que les hommes y eussent trouvé, avec soulagement, l'occasion souhaitée d'oublier les haines, et de fraterniser dans un effort totalement désintéressé.

La France étant la puissance invitante, le français avait été choisi comme langue de travail.

Mais pour rendre les relations plus faciles, le Japon avait installé à EPI 2 une Traductrice universelle à ondes courtes. Elle traduisait immédiatement les discours et dialogues qui lui étaient transmis, et émettait la traduction en 17 langues sur 17 longueurs d'ondes différentes. Chaque savant, chaque chef d'équipe et technicien important, avait reçu un récepteur adhésif, pas plus grand qu'un pois, à la longueur d'onde de sa langue maternelle, qu'il gardait en permanence dans l'oreille, et un émetteur-épingle qu'il portait agrafé sur la poitrine ou sur l'épaule. Un manipulateur de poche, plat comme une pièce de monnaie, lui permettait de s'isoler du brouhaha des mille conversations dont les 17 traductions se mélangeaient dans l'éther comme un plat de spaghetti de Babel, et de ne recevoir que le dialogue auquel il prenait part.

(...)

D'une voix chantante, un peu monotone, Léonova fit le point des travaux, et la Traductrice se mit à chuchoter dans toutes les oreilles, en dix-sept langues différentes. Léonova se tut, resta un instant rêveuse, et reprit.
- Je ne sais pas ce que vous suggère la vue de cette sphère, mais moi...elle me fait penser à une graine. Au printemps, la graine devait germer. (...) Mais le printemps n'est pas venu...Et l'hiver dure depuis 900 000 ans...Pourtant, je ne veux pas, je ne peux pas croire que la graine soit morte...

(...)

Debout sur l'escalier devant la porte de l'Œuf, Hoover donnait des informations sur les travaux de son équipe. Dans la salle de Conférences, les journalistes regardaient sur le grand écran, et prenaient des notes.
- Nous l'avons percée ! dit Hoover, Voici le trou...
Son gros pouce se posa sur la porte près d'un orifice noir dans lequel il aurait juste pu s'enfoncer.
- Il n'y a eu de mouvement d'air ni dans un sens ni dans l'autre. L'équilibre des pressions interne et externe ne peut pas être l'effet du hasard. Il y a quelque part un dispositif qui connaît la pression externe et agit sur la pression interne. Où est-il ? Comment fonctionne-t-il ? Vous aimeriez bien le savoir ? Moi aussi...
Rochefoux parla dans le micro de la table du Conseil.
- Quelle est l'épaisseur de la porte ?
- Cent quatre-vingt-douze millimètres, composés de couches alternées de métal et d'une autre matière qui semble être un isolant thermique. Il y a au moins une cinquantaine de couches. C'est un vrai feuilleté...Nous allons mesurer la température intérieure.
Un technicien introduisit dans l'orifice un long tube métallique qui se terminait, à l'extérieur, par un cadran. Hoover jeta un coup d'oil sur ce dernier, eut l'air brusquement intéressé et ne le quitta plus des yeux.
- Eh bien, mes enfants !...Ça descend ! Ça descend !...Encore...Encore... Nous sommes à moins 80...moins 100...120...Il cessa d'énumérer les chiffres et se mit à siffler d'étonnement. La Traductrice siffla sur ses dix-sept écouteurs.
- Moins 180 degrés centigrades ! dit l'image de Hoover en gros plan. C'est presque la température de l'air liquide !

(...)

- Bon, dit Deville. Vous avez fait un trou dans la glace, vous avez trouvé une graine. Vous avez fait un trou dans la graine, vous avez trouvé un œuf. Aujourd'hui, qu'est-ce que vous allez trouver, à votre idée ?
Hoover fit face avec un sourire charmant sur son gros visage.
- Nuts *? dit-il.*
Ce que la traductrice, après un millionième de seconde d'hésitation, traduisit dans les micros français par :
- Des noix ?
- Des clous ?
- Il ne faut pas trop demander à un cerveau automatique. Pour conserver l'image ronde, un cerveau d'homme aurait peut-être traduit par « des prunes » ?

René Barjavel, *La nuit des temps*, 1968

Contexte

La thèse a été effectuée dans l'équipe GETALP du laboratoire LIG (ex CLIPS-IMAG), en convention CIFRE avec la société Prosodie.

Laboratoire LIG et équipe GETALP

Début 2007, cinq laboratoires d'informatique grenoblois ont fusionné pour former le Laboratoire d'Informatique de Grenoble (LIG). À cette occasion, les équipes GETA (Groupe d'Étude pour la Traduction Automatique) et GEOD (Groupe d'Étude sur l'Oral et le Dialogue) de l'ancien laboratoire CLIPS (Communication Langagière et Interaction Personne-Système) se sont réunies en une équipe unique au sein du LIG : l'équipe GETALP (Groupe d'Étude pour la Traduction Automatique et le Traitement Automatisé des Langues et de la Parole).

L'équipe GETALP est organisée autour de cinq thèmes de recherche principaux :

1. Traduction Automatique (TA) et Automatisée (TAO).
2. Reconnaissance automatique de la parole, des locuteurs et des sons.
3. Ressources lexicales et corpus (logiciel et contenu).
4. Dialogue, communication et émotions.
5. Langages de programmation et environnements spécialisés pour le TALN.

Les activités de ces thèmes de recherche partagent cinq défis :

1. Rendre l'informatique multilingue et « ubilingue ».
2. Informatiser les langues peu dotées et peu écrites en adaptant des ressources existantes.
3. Rendre la communication langagière multimodale (texte, parole, geste).
4. Trouver et implémenter des méthodes et outils d'évaluation liés à la tâche.

Avant la thèse, l'équipe avait notamment déjà travaillé sur la traduction de dialogues oraux finalisés, dans le cadre des projets C-STAR II (p. 57) et Nespole ! (p. 61).

Société Prosodie

Prosodie est une société créée en 1986, et spécialisée dans la conception et l'hébergement de solutions liées à la relation client multicanal et au commerce électronique. Elle dispose également d'une offre d'infogérance d'applications Web et de systèmes d'information. Les clients de Prosodie sont essentiellement des grands comptes et des administrations couvrant divers secteurs d'activité. Présent en France, en Espagne, aux États-Unis et au Canada, le groupe Prosodie a réalisé en 2007 un chiffre d'affaires consolidé de 172 200 000 euros.

Prosodie s'est notamment illustrée en 2003, par la mise en œuvre d'un procédé de reconnaissance vocale assistée par opérateurs (RVAO), dans le cadre de l'aiguillage d'appels téléphoniques. À son arrivée sur un service de RVAO, un serveur vocal invite l'appelant à nommer le service auquel il souhaite accéder, l'appel est alors dirigé vers le service approprié. L'identification du service se fait par reconnaissance vocale, mais lorsque cette dernière ne parvient pas à un taux de confiance suffisant, c'est un opérateur qui prend le relai pour diriger l'appel vers le bon service.

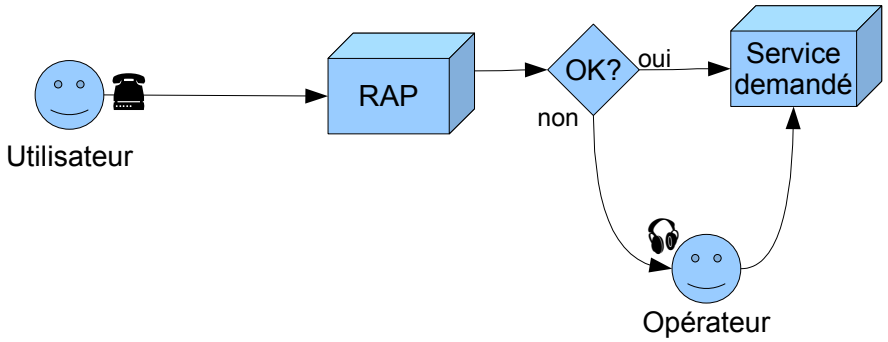


Figure 1 – Principe de la Reconnaissance Vocale Assistée par Opérateur.

Ce système traite plus de 2 millions d'appels par jour, dans le cadre de plus d'une centaine de services vocaux, tels que la météo en ligne ou le télé-achat.

Vers la thèse

Or, en 2000, la société a été sensibilisée à la traduction de dialogues oraux finalisés lors d'une présentation de C-STAR II, terminé l'année précédente, effectuée par Christian Boitet et Hervé Blanchon. À la suite de cette présentation, des discussions ont commencé avec Alain Bernard, alors PDG de Prosodie, en vue d'un projet de système d'aide au dialogue en langue seconde, pour téléphone logiciel[1], visant à spécifier un produit et concevoir un prototype pour l'anglais.

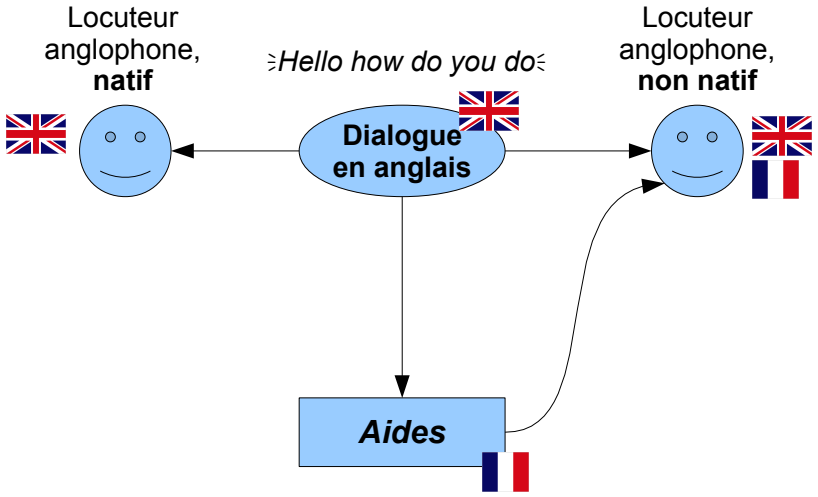


Figure 2 – Vue du projet initial : discussion en anglais, incluant un locuteur de langue native française.

[1] Aussi appelé *softphone* : logiciel de téléphonie sur Internet, comme par exemple Skype.

Introduction

Comme on peut le voir à la lumière des deux textes en exergue de cette thèse, la question du traitement de la communication ne laisse pas indifférent. À la vision d'une humanité réduite à l'état de machinerie communicante, frémissant de bourdonnements vides de sens, s'oppose l'espoir d'un dépassement des limites de la communication naturelle, à commencer par celle de la barrière linguistique.

Cela fait déjà plusieurs années que le TALN a investi la problématique du dialogue multilingue. La traduction de dialogues en temps réel est aujourd'hui assez bien maîtrisée, à condition de se limiter à un sous-domaine d'application restreint, et à des énoncés formellement « corrects » et relativement neutres. Dans ce cas, en effet, l'intelligence des locuteurs et le caractère fermé de la tâche suffisent généralement pour faire progresser le dialogue. Ainsi, le projet C-STAR (p. 57) a donné des résultats intéressants pour le dialogue touristique ; le projet Nespole ! (p. 61), pour le dialogue touristique et financier ; ou encore MASTOR, un outil d'IBM, pour le tourisme, la médecine et la sécurité. Mais dès que l'on cherche à étendre le domaine d'application de tels systèmes, le taux d'erreur devient problématique.

Les systèmes de traduction automatique de la parole se basent généralement sur deux procédés : soit une traduction automatique statistique «directe», soit un traitement en trois étapes, extraction du sens, génération d'un énoncé en langue cible, puis synthèse vocale. Des erreurs peuvent se produire au niveau de la reconnaissance de la parole, comme de la traduction.

En outre, ces outils ont en commun d'être élaborés selon une perspective que l'on pourrait qualifier de « médiation forte » : l'outil vient s'intercaler entre des locuteurs que l'on suppose incapables de communiquer seuls. Cette perspective se justifie dans certains cas, mais s'avère souvent contraignante, à plusieurs égards. D'une part, elle exige pour être utilisable une chaîne de linguiciels homogènes et de très grande qualité, qui restent aujourd'hui difficiles à obtenir en dehors de quelques sous-domaines restreints. D'autre part, elle fait l'hypothèse que les utilisateurs de ces outils n'ont pas d'autre moyen de communiquer ; si les locuteurs partagent une langue commune, même avec un niveau assez faible, les compétences linguistiques des utilisateurs, leurs stratégies de communication employant cette langue commune ne sont pas prises en compte. Enfin, une communication totalement médiée par la machine, même de qualité, vient parasiter la dimension phatique de la communication, et peut donc ne pas être souhaitée par certains utilisateurs.

Afin de parvenir à un système plus satisfaisant dans un contexte où les locuteurs maîtrisent, plus ou moins bien, une langue commune, il est nécessaire d'envisager d'autres approches. L'une d'elles consiste à imiter un locuteur humain, qui demande à son interlocuteur de répéter, de préciser certains points, lorsqu'il comprend mal. Dans un contexte applicatif, cela correspond à un procédé de désambiguïsation interactive. Cette dernière intervient à l'issue du traitement automatique, et permet à un humain de guider le système dans ses choix les plus critiques.

Cette piste a déjà été explorée avec succès pour l'écrit documentaire dans le cadre du projet LIDIA.Cette méthodologie donne une traduction de très bonne qualité, d'autant que les informations de désambiguïsation peuvent servir pour des traductions dans plusieurs langues, et est utilisable dans le cadre de traduction de documents, mais risque d'être perçue comme fastidieuse dans le cadre d'un dialogue. Or, nous nous plaçons ici dans le cadre du dialogue, et l'on doit tenir compte de cette nouvelle dimension. C'est à dire qu'il faut s'intéresser, non seulement comme on le fait classiquement au sens propositionnel de chaque énoncé considéré isolément, mais aussi traiter les niveaux plus profonds des intentions et des émotions, qui ne sont pas pris en compte pour le traitement de l'écrit documentaire, dans le contexte interactif d'un dialogue en temps réel.

La thèse a pour objectif de spécifier et de prototyper des outils afin de permettre aux locuteurs natifs et non natifs[2] de résoudre leurs problèmes linguistiques.

Dans un premier temps, nous nous sommes orientés vers un système d'aide à la conversation orale, un « tradphone ». Une étude des dialogues en langue étrangère a été menée, dans le sens de la caractérisation de la langue non native et de la mise en évidence les problèmes rencontrés par les locuteurs. Diverses fonctionnalités intéressantes ont été envisagées, mais il s'est avéré que le vecteur d'entrée obligé du système, la reconnaissance vocale, était aussi son principal point faible, et qu'une amélioration très significative de cette dernière dans un contexte téléphonique et non natif était largement au-delà du cadre de la thèse. C'est pourquoi, sans exclure totalement une entrée vocale, nous avons préféré privilégier l'entrée écrite et passer à une problématique de tchat[3] en langue étrangère.

Dans le domaine du tchat écrit, beaucoup restait néanmoins à faire, tant au niveau de l'étude de la langue que de son outillage. Nous nous sommes donc employé à décrire précisément la « langue » du tchat, qui s'avère paradoxalement très proche de la graphie des langues anciennes, et d'appréhender la motivation des comportements langagiers observés, mettant en évidence certaines différences significatives vis à vis des autres types d'écrits propres au Web. Nous avons ensuite évalué dans quelle mesure les caractéristiques de la langue orale non native étaient applicables à la langue du tchat. Enfin, nous avons établi et implémenté, sous forme d'application web, une palette d'outils destinés à aider les interlocuteurs à surmonter les obstacles

[2]Voir p. 41 pour une définition de ces termes.
[3]Voir p. 12 pour une justification de ce choix orthographique.

de la communication en langue non native (en particulier une désambiguïsation participative), et déterminé comment pondérer, en fonction de la situation, de la finalité du dialogue et du niveau du locuteur, divers instruments de mesure linguistique : notamment mesure de l'intelligibilité et de la prototypicalité du discours.

Dans une première partie, nous présenterons, dans ses lignes générales, le dialogue, d'abord dans sa modalité la plus étudiée : le dialogue oral spontané. Nous nous pencherons alors sur le dialogue écrit spontané, dont le meilleur représentant, depuis l'apparition des réseaux informatiques, est médié par la machine : c'est le tchat.

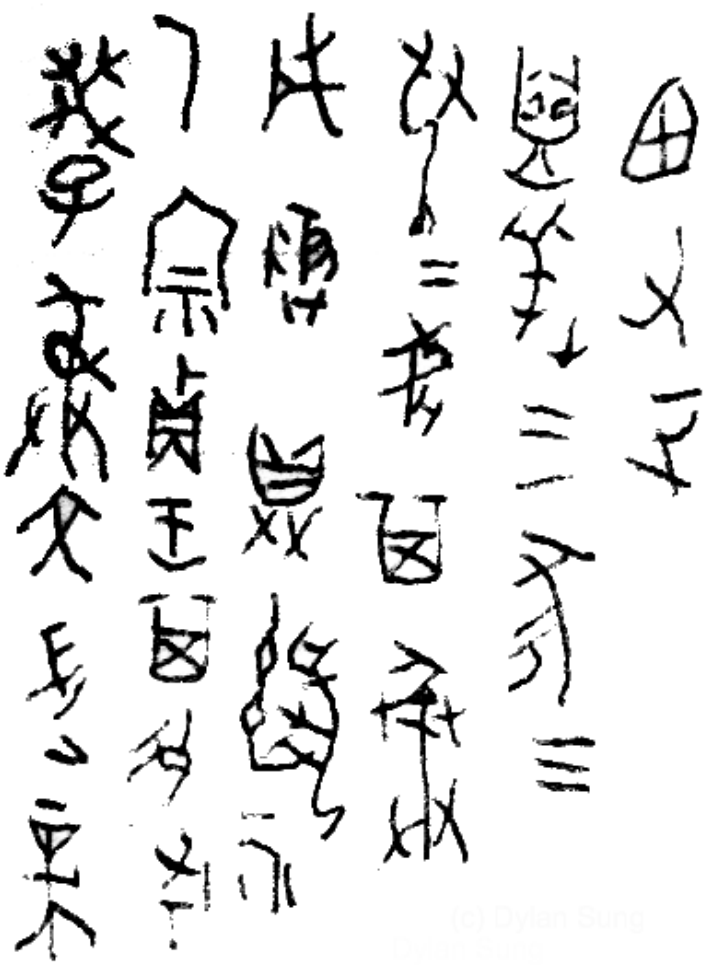
Nous verrons ensuite quels problèmes se posent aux locuteurs dans le cadre de dialogues en langue seconde, quelles sont leurs stratégies de résolution dans un cadre « naturel », sans aides informatiques, et ce que cela entraîne comme caractéristiques problématiques si l'on souhaite informatiser ce dialogue. Nous verrons comment les outils actuels tentent de répondre aux problèmes ainsi mis en évidence, tant du point de vue du traitement informatique du langage, que du point de vue de la présentation de ces traitements aux utilisateurs.

En conséquence, nous proposerons une approche plus adaptée à ce contexte, basée sur les observations effectuées dans les deux parties précédentes, tant du point de vue des fonctionnalités à implémenter, que de leur présentation aux utilisateurs. Nous nous attacherons en particulier, à organiser ces fonctionnalités de façon non bloquante pour les utilisateurs, de façon à ce que le dialogue puisse progresser de façon normale, et à ce que les aides apportées répondent aux problèmes des utilisateurs, mais aussi que leur utilisation suive les stratégies de résolution mises en œuvre naturellement par ces derniers.

Enfin, nous présenterons la mise en œuvre de cette approche dans le cadre d'un prototype, sous forme d'application web, en vue d'une évaluation.

Chapitre 1

Le dialogue parlé et écrit : état de l'art et apports



Une des premières inscriptions en chinois archaïque sur des ossements. Dynastie Zhou, début du premier millénaire avant JC.

1 Le dialogue

Pour atteindre l'objectif de notre travail, concernant l'aide au dialogue en langue seconde, il est utile de s'intéresser dans un premier temps au dialogue proprement dit. Cette première partie va s'y attacher, mais sans approfondir cette étude plus que ce qui est nécessaire pour atteindre notre but.

L'intérêt pour le dialogue « naturel », c'est à dire spontané et non pas littéraire, est relativement récent. Ce type de dialogue fut longtemps perçu comme quelque chose de trivial, indigne d'intérêt scientifique, alors même que la primauté de la langue naturelle et de sa réalisation dans le cadre du dialogue étaient reconnus par Saussure dans ses travaux précurseurs dès les années 1910 [Sau16]. Il faudra attendre l'émergence du structuralisme dans les années 1960 pour que le dialogue apparaisse en tant que véritable objet d'étude scientifique (voir par exemple Jakobson [Jak63]). Cet intérêt pour le dialogue spontané est toujours contesté aujourd'hui, notamment en France [Cah99]. Le dialogue reste souvent assimilé à une forme dégradée de discours [KO05], puisqu'après tout « on parle toujours pour quelqu'un ». Néanmoins, cette distinction nous apparaît essentielle si l'on souhaite réaliser des outils vraiment adaptés à une utilisation en dialogue.

1.1 Présentation et outils de description

Dans son ouvrage « Le discours en interaction » [KO05], Catherine Kerbrat-Orecchioni pose l'interactivité comme caractéristique centrale du dialogue par rapport au discours, en particulier la possibilité d'intervention dans le processus de construction des énoncés de l'interlocuteur (coconstruction des énoncés). Cela suppose une grande réactivité des interlocuteurs, qui exclut notamment le « dialogue » épistolaire ; selon cette acception stricte, elle pose le tchat comme seule forme de dialogue écrit.

L'unité de base du dialogue est le *tour de parole* [DS95], c'est à dire une prise de parole idéalement ininterrompue et exclusive. Toutefois, en dehors des constructions artificielles telles que le dialogue littéraire ou théâtral, cette définition n'est pas stricte ; il ne faut pas s'attendre à pouvoir décomposer un dialogue en une stricte succession de tours de parole, l'interlocuteur pouvant intervenir de diverses manières dans la construction des énoncés en dehors de son tour [BB00]. On observe donc à divers degrés des recouvrements de tours de parole, plusieurs locuteurs s'exprimant simultanément sans que cela ne soit perçu comme nuisant à la conversation, bien au contraire (cf. table 1.1).

(1)	Loc1	si vous êtes par exemple euh directeur d'une école - communale
(2)	Loc2	ça gagne plus
(3)	Loc1	vous allez gagner 400 francs de plus

Table 1.1 – Intervention dans le tour de parole de l'interlocuteur [BB00].

L'allocation des tours de parole est un processus collaboratif au cours duquel les locuteurs négocient leurs prises de parole. Par exemple, par la prosodie (fréquence fondamentale descendante) ou l'emploi de formules adaptées (« voilà »), le locuteur peut marquer la fin de son tour.

Les tours de parole s'organisent souvent par paires : question/réponse, échange de salutations, offre/(acceptation-refus), clôture, etc. À une action accomplie dans un premier énoncé doit répondre une action appropriée de l'interlocuteur. On parle alors de *paires adjacentes*. Souvent, l'action attendue de l'interlocuteur est contrainte (compliment, accusation, reproche, etc.) : certaines répliques auront tendance à être préférées à d'autres. Cela est d'ailleurs anticipé par le premier locuteur, qui a tendance à formuler son énoncé en tenant compte de la réponse préférée attendue.

La structure d'un dialogue, ou au moins de certaines portions d'un dialogue, est prédéterminée selon leur finalité et la nature des intervenants. On parle de script pour désigner ces patrons archétypaux [KO05] (cf. table 1.2). Les scripts varient

(1)	Loc1	salutation
(2)	Loc2	salutation
(3)	Loc1	exposé du problème
(4)	Loc2	accusé de réception
(5)	Loc1	question
(6)	Loc2	question de clarification
(7)	Loc1	réponse
(8)	Loc1	solution
(9)	Loc1	remerciement

Table 1.2 – Script de consultation d'un expert. La paire 6-7 est répétée tant qu'il reste des points à éclaircir. Remarquer aussi l'organisation de la plupart des tours de parole en paires adjacentes.

parfois suivant la culture. Ainsi les scripts d'achat client/commerçant comportent une phase de marchandage dans certaines cultures, alors qu'ailleurs cette phase ne sera pas présente par défaut. Pour demander un service à un subordonné, un anglophone s'entourera de précautions oratoires (« *would you mind...* »), tandis qu'un francophone sera plus direct [KO05].

1.2 Accidents et bruits

La *théorie de la communication* [WES63] se base sur un modèle mécaniste simple initialement développé pour la téléphonie (cf. figure 1.1). Le dialogue est considéré comme un échange d'informations. Ce modèle mécaniste permet d'analyser les problèmes qui surviennent lors de la communication. Ces problèmes sont de trois ordres [Bou93] : techniques, sémantiques et opérationnels.

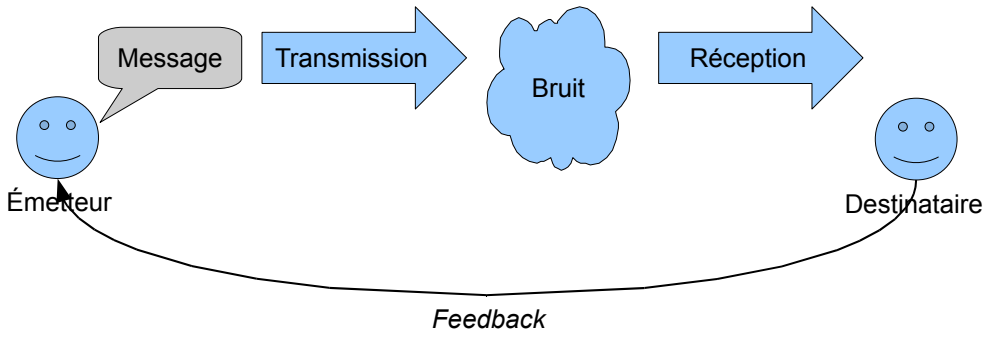


Figure 1.1 – Modèle Shannon–Weaver du canal bruité.

Les problèmes techniques concernent l'exactitude du transfert des symboles. Dans le cas de la communication parlée, il peut s'agir des phonèmes, dont la perception peut être brouillée s'ils sont mêlés à d'autres sons, mais aussi de l'intonation, et même des expressions faciales et des la gestes, qui participent à l'expression du locuteur. À un plus haut degré, le décodage lexical peut aussi poser problème, par exemple en cas d'homophonie.

(1)	Loc1	Le goût c'est lié au nez
(2)	Loc2	Dis-donc qu'est-ce que t'es chauvine ! Les lyonnais n'ont quand même pas le monopole du goût !

Table 1.3 – Problème technique : homophonie [KO05] : au premier tour de parole, le locuteur Loc2 a entendu *lyonnais* à la place de *lié au nez*.

Les problèmes sémantiques concernent l'identité entre l'interprétation du récepteur et l'intention de l'émetteur, et peuvent être causés par des énoncés ambigus, sous-déterminés au regard des présupposés des locuteurs. Non détectés par les locuteurs, ils peuvent donner lieu à des malentendus plus ou moins graves, qui ne peuvent souvent être détectés et corrigés que tardivement, lorsqu'un faisceau de présomptions conduit à des contradictions évidentes, ou en cas de contestation de l'interlocuteur [CN06].

(1)	Loc1	d'accord là vous étiez aux États-Unis
(2)	Loc2	pas aux États-Unis euh – au boulevard des États-Unis
(3)	Loc1	ah à Lyon

Table 1.4 – Dialogue recruteur/candidat ; ce dernier a indiqué sur son CV avoir travaillé « aux États-Unis » [KO05].

Les problèmes d'efficacité (« problèmes pragmatiques », pourrait-on dire) concernent le succès avec lequel la signification transmise jusqu'au destinataire, provoque chez lui la conduite désirée.

Un problème de classe inférieure (*technique* par exemple : un mot confondu avec un autre de prononciation proche) pourra provoquer des problèmes en cascade aux niveaux suivants (incompréhension, voire malentendu, pouvant lui-même provoquer une conduite inappropriée).

Cette théorie voit dans la redondance un moyen d'éviter ces problèmes : répétitions de la phrase, des mots importants, gestes, intonation, en fonction du retour (*feedback*) de l'interlocuteur.

Lorsque l'on considère l'écrit comme étant la norme, le dialogue oral semble particulièrement bruité. Or, beaucoup de signes produits par l'interlocuteur, qui pourraient apparaître comme des bruits inutiles voire parasites, peuvent en fait s'avérer utiles pour le dialogue.

> *Nombre de faits que l'on a coutume de considérer dans le discours oral, se référant à la norme du discours écrit, comme des ratés et des « bruits », apparaissent au contraire, dès lors qu'on les prend pour ce qu'ils sont, c'est à dire des phénomènes de nature interactive, comme éminemment fonctionnels. Au lieu de démontrer le caractères défectueux des sujets parlants, de tels phénomènes constituent autant de manifestations de leur capacité à construire des énoncés efficaces interactivement.*

Catherine Kerbrat-Orecchioni, *Le discours en interaction* [KO05], page 47.

Tout d'abord, les signes simples marquant l'attention (bruits de gorge, sifflements, mouvements de la tête, regard, expression faciale, etc.), agissent comme *signaux d'écoute* [KO05]. Ils permettent de signaler le bon fonctionnement du canal de communication, et l'attention que l'on porte aux propos de l'interlocuteur. Les signes d'acquiescement plus marqués (*continuateurs* [KO05] vocaux ou visuels) tendent à montrer que l'interlocuteur interprète correctement le message, ou du moins qu'il n'y détecte pas d'incohérence. Le fait d'intervenir dans le tour de parole d'autrui (cf. exemple 1.1), de répéter ce qu'il vient de dire, peut être utilisé afin de montrer l'identité de vue entre les locuteurs. Enfin, certains signes servent à structurer le dialogue. Dans l'exemple 1.5, « euh ben voilà » est utilisé pour indiquer l'ouverture d'une paire question/réponse. Le « oui » du locuteur 1 marque l'écoute de ce dernier mais aussi son acceptation de cette ouverture.

(1)	Loc1	bonjour monsieur
(2)	Loc2	bonjour madame
(3)	Loc1	je vous écoute
(4)	Loc2	euh ben voilà - je voudrais savoir (...)
(5)	Loc1	oui

Table 1.5 – Structuration du dialogue [KO05].

De son côté, le locuteur peut « demander » une rétroaction de son interlocuteur [KO05], par exemple, une pause prolongée (parfois remplacée par un « euh ») [Goo81], ou une hésitation [Owe81]. Bien sûr, toutes les pauses et hésitations ne sont pas fonctionnelles, et nombre d'entre elles sont dues à la quasi-concomitance de la planification et de la production propres à l'oral, les erreurs et hésitations dans la planification se trouvant immédiatement exposés dans la parole. Mais ces « demandes » apparaissent tout de même loin d'être anecdotiques. Ainsi, pour ce qui est des pauses, 35% d'entre elles assumeraient clairement un rôle fonctionnel [Goo81].

1.3 Coconstruction des énoncés

La construction interactive des énoncés ne se limite pas au niveau technique de la théorie de la communication. Au niveau sémantique, l'interlocuteur peut venir en aide au locuteur, apporter des précisions dans ses énoncés (exemple 1.6). Il peut aussi réparer des erreurs commises par ce dernier (table 1.7).

(1)	Loc1	alors selon les variétés, ça peut monter à quinze mètres de haut et avoir six mètres d'étalement
(2)	Loc2	oui euh – quelque chose comme huit à dix mètres
(3)	Loc1	oui - huit à dix mètres

Table 1.6 – Confirmation. Consultation d'expert à la radio [KO05].

(1)	Loc2	(...) mais vous en avez encore la possibilité puisque cet arbre est encore jeune d'après ce que j'ai compris — c'est de le déplacer
(2)	Loc1	il a une dizaine d'années
(3)	Loc2	ah une dizaine d'années - alors ça va être dur il vaut mieux ne pas le déplacer

Table 1.7 – Hétéroréparation. Consultation d'expert à la radio [KO05].

Dans le cas d'un malentendu, la réparation peut survenir assez loin dans le dialogue (exemple d'auto-réparation tardive : table 1.8), sombrant parfois dans le dialogue de sourds.

On remarquera que tous ces phénomènes de réparation s'accompagnent souvent d'excuses, de rires, de marques de considération (« ah oui », « c'est vrai », « bien sûr », etc.) et d'autres marqueurs/testeurs de bon fonctionnement du canal de communication. Ces marqueurs/testeurs sont donc plus fréquents lors de situation problématiques. On peut penser qu'ils sont alors un moyen de renforcer le « contact ». Loin d'être des bruits parasites, ils joueraient donc un rôle facilitateur dans le dialogue.

(1)	Loc1	alors à Lyon ça c'est bien passé votre week-end c'est vraiment bête que j'aie pas pu être là
(2)	Loc2	ben : : juste en arrivant on s'est tapé un super bouchon
(3)	Loc1	un vendredi soir il devait y avoir de l'ambiance
(4)	Loc2	ça oui
(5)	Loc1	c'était dans le tunnel ?
(6)	Loc2	dans le tunnel ? – ah : : : *(rires)* mais non un bouchon - lyonnais quoi - on s'est tapé un p'tit restau

Table 1.8 – Malentendu autoréparé tardivement [KO05].

1.4 Empathie et congruence

Afin de bien considérer le rôle de cette fonction de « contact », nous avons besoin de sortir du cadre de la théorie de la communication. En effet, cette théorie aborde la communication interpersonnelle en tant qu'outil de transmission de l'information, mais néglige les aspects psychologiques. La fonction phatique du langage, telle que la définit Jakobson [Jak63], ne s'écarte pas non plus beaucoup de l'idée d'un simple rôle de vérification du canal. Mais l'omniprésence de ces marqueurs, leur multiplication dans le cas de problèmes où le canal n'est pas en cause, peuvent-elles être entièrement réduites à ce rôle mécanique ? La psychologie humaniste apporte un point de vue permettant de compléter l'approche mécaniste.

Le psychologue Carl Rogers [Rog51] a longuement étudié le rôle de la parole d'un point de vue thérapeutique. Pour lui, la parole n'a pas seulement pour fonction de faire parvenir une information à l'interlocuteur. Dans le cadre de la thérapie, la parole tient aussi pour le locuteur un rôle d'expression de soi, d'objectivation. Un locuteur aura tendance à s'identifier à ses propres productions, tant en ce qui concerne les propos (sens, mais aussi formulation, dimension « poétique ») qu'en ce qui concerne la voix. Pour l'interlocuteur, la parole qui lui est destinée est un acte de représentation, mais aussi, du fait même de l'effort de représentation dont il est destinataire, un acte de reconnaissance de son état de sujet.

Il a été proposé d'étendre cette notion à l'ensemble de la communication interpersonnelle [MP00], à travers deux conditions nécessaires à une bonne communication entre deux personnes : l'empathie et la congruence. L'empathie concerne la capacité à émettre et recevoir des signaux qui renseignent sur l'état mental et les affects de l'individu. Elle pose le locuteur et son interlocuteur en tant que sujets. Quand les paroles du locuteur sont en accord avec ces pensées, on parle de congruence. L'interlocuteur sent qu'il y a accord entre ce qui est dit et ce qui est pensé. Dans le cas contraire il n'y a pas congruence : l'interlocuteur repère la situation incongrue et ne fait pas confiance.

Au cours du dialogue, les locuteurs cherchent à maintenir à la fois l'empathie et la congruence, à travers de nombreux signes, verbaux ou non. En cas de situation

troublée, comme par exemple l'émergence d'un malentendu, ou un désaccord chez des personnes recherchant un compromis, ces signes vont se multiplier, témoignant d'un effort d'empathie et de congruence accru.

Conclusion

Deux points nous semblent importants à retenir. D'une part nous considérerons le dialogue comme une activité collaborative, et non comme une stricte succession de tours de paroles indépendants et préparés à l'avance. Les tours de parole sont l'objet d'un débat, d'une collaboration, non seulement au niveau de leur répartition, mais aussi en ce qui concerne leur construction. D'autre part, les locuteurs produisent continuellement des signes de bonne réception, mais aussi d'empathie, qu'une approche fondée sur la norme peut amener à considérer comme parasites, alors qu'ils sont indispensables. En effet, confrontés à une difficulté, les locuteurs vont multiplier ces signes.

Pour un outil d'aide au dialogue en langue seconde, nous pourrons donc nous appuyer sur cette capacité de collaboration et d'entraide entre utilisateurs. Par contre, il faudra prendre garde à conserver une dimension « empathique », c'est à dire des signaux ne faisant pas explicitement partie de la norme. Cela pose problème dans le cadre d'une approche de type reconnaissance vocale - traduction - synthèse vocale, qui aura tendance à filtrer les marqueurs de l'empathie.

2 Le dialogue écrit : le tchat

2.1 Qu'est-ce que le tchat ?

2.1.1 Nom et graphie

Le tchat est souvent associé avec l'idée d'une langue assez floue, peu normée, et cela se retrouve jusque dans la façon de le nommer. Plusieurs dénominations ont cours en français :

chat : prononcé /tʃat/, emprunté à l'anglais *to chat* « discuter » ;

tchat : une « francisation » du terme précédent, sur le modèle du terme d'argot pied-noir « tchatcher » (de l'espagnol *chacharear*, « bavarder »), passé en français populaire ;

t'chat : une sorte de compromis entre les deux formes précédentes ;

clavardage : néologisme recommandé par l'Office Québécois de la Langue Française en 1997 [OQL] ;

causette : néologisme recommandé par la Commission Générale de Terminologie et de Néologie (France) de 1999 à 2006 [JO9] ;

dialogue en ligne : néologisme recommandé par la Commission générale de terminologie et de néologie, en lieu et place de « causette », depuis 2006 [JO0].

Le terme « causette » n'est pratiquement jamais rencontré dans le sens de « tchat », y compris sur les sites officiels qui privilégient la graphie « chat » [FOG]. Quant au terme « dialogue en ligne », c'est plus une description qu'un terme à proprement parler, ce qui n'est pas très satisfaisant. Une rapide recherche sur les formes verbales dérivées[1] des quatre termes restants, permet d'estimer grossièrement leur fréquence respective.

Graphie(s) recherchée(s)	Nombre de pages	Pourcentage
chater + chatter	664 000	41
tchater + tchatter	878 000	55
t'chater + t'chatter	17 930	1
clavarder	40 200	3
total	1 600 130	100

Table 1.9 – Nombre de pages francophones référencées par Google.com en novembre 2008 pour diverses graphies du verbe « tchater ».

Nous avons choisi de privilégier la graphie « tchat », la plus courante.

2.1.2 Définitions

Étant donné le nombre toujours croissant de modes de communication disponibles sur Internet, il est utile de rappeler ce que nous entendons précisément par le terme « tchat ». **Par le terme « tchat », on entend un dialogue écrit, en temps réel et médié par l'informatique.** Le tchat présente de nombreuses similarités avec le dialogue oral ; et c'est sans doute le moyen de communication écrit qui s'en rapproche le plus. Il en est nettement plus proche que d'autre outils, tels que les forums ou les e-mails, que le caractère différé assimile plus au dialogue épistolaire qu'à la conversation parlée. De plus, le tchat est volatil. Le contenu d'une session de tchat, à l'instar d'une conversation orale, n'a pas vocation à être enregistré ; lorsqu'ils se connectent, les utilisateurs ne savent pas ce qui a été dit précédemment, et ne sauront pas plus ce qui se dira après leur déconnexion.

On associe souvent tchat et texto (ou SMS[2]). Mais à la différence du texto, les outils de tchat proposent aux utilisateurs des espaces communs, qui rendent aisée la communication entre de nombreuses personnes, là où le texto peut difficilement mettre en relation plus de deux personnes. De plus, il faut garder à l'esprit que les textos sont facturés au message, et composés à l'aide d'un clavier numérique peu adapté, ce qui incite l'utilisateur à être synthétique et à envoyer aussi peu de caractères que possible, alors qu'au contraire le tchat, moins contraignant de ce point de vue, autorise toutes les digressions. Ces différences de contexte de production ont une grande influence sur le « langage » produit : on verra par la suite qu'il existe des différences significatives entre les productions linguistiques du tchat et celles du texto.

[1] Afin de contourner le problème d'homographie posé par *chat*.
[2] *Short Message Service*

On distingue parfois tchat et messagerie instantanée. Le terme tchat désigne alors une discussion publique, reposant presque exclusivement sur le protocole ouvert IRC[3] ; nous utiliserons cet acronyme « IRC » par la suite pour désigner cette définition restrictive du tchat) ; conversation ouverte entre de nombreux interlocuteurs inconnus, prenant place sur des canaux ou salons de discussion. Dans cette distinction, le terme messagerie instantanée désigne une discussion privée, reposant sur une grande variété de protocoles fermés tels que MSNP[4] (cf. figure 1.2), YM[5], ou ouverts, principalement XMPP[6] ; conversations privées impliquant seulement deux personnes se connaissant déjà. En pratique, cette distinction est très floue : il est possible de rendre privé un canal IRC, seule une liste de personnes connues pouvant s'y connecter ; à l'inverse, il est possible sur les outils de messagerie instantanée de mener des discussions avec de nombreux interlocuteurs, et n'importe lequel de ces interlocuteurs peut inviter dans la conversation l'un de ses contacts, même inconnu des autres participants. Ces deux termes sont même bien souvent interchangeables. L'autre différence majeure entre la messagerie instantanée et IRC tient au nombre

Figure 1.2 – *Windows Live Messenger*, le client officiel du réseau de messagerie instantanée de Microsoft (source : *Wikimedia Commons*).

d'utilisateurs de ces deux systèmes. Sur les principaux réseaux IRC, IrcNet, EFNet,

[3] *Internet Relay Chat*
[4] *MicroSoft Network Protocol*
[5] *Yhaoo ! Messenger*
[6] *eXtensible Messaging and Presence Protocol*

DalNet et UnderNet, le nombre d'utilisateurs dépassait rarement 150 000 en 2001, et avait tendance à baisser [LT01b], tandis que le nombre d'utilisateurs réguliers des services de messagerie instantanée était estimé à 150 000 000 cette même année, en forte croissance [Mag]. Certes, le nombre d'utilisateurs réguliers d'IRC doit largement dépasser ces 150 000 connexions simultanées, mais le rapport reste écrasant ; de fait, les logiciels de messagerie instantanée, tels que MSN, AIM, ICQ, YM ou encore Gaim/Pidgin, ont aujourd'hui bien plus de notoriété que mIRC, le plus connu des clients IRC. Finalement, nous ne voyons qu'une seule différence claire entre ces deux modes de communication : l'accès à IRC se fait en sélectionnant un canal de discussion, sans savoir *a priori* qui s'y trouve connecté (le canal peut même être vide) ; alors qu'en messagerie instantanée l'accès se fait en choisissant les personnes avec qui l'on souhaite communiquer (généralement dans une liste de contacts, mais on peut aussi utiliser simplement l'adresse de cette personne).

Cette distinction nous semble intéressante d'un point de vue sociologique et sociolinguistique, mais peu pertinente d'un point de vue strictement linguistique ou ergonomique, c'est pourquoi dans la plupart des cas nous ne séparons pas ces deux notions, et parlons tout simplement de « tchat » pour désigner indifféremment « IRC » et « messagerie instantanée ».

2.1.3 Aperçu

Bien qu'il existe de nombreux logiciels et réseaux de tchat, on peut relever une architecture constante (du moins pour les systèmes utilisés aujourd'hui, voir l'historique). Le tchat s'appuie sur une architecture de type client/serveur ; c'est à dire que les utilisateurs ne communiquent pas directement entre eux, mais par l'intermédiaire d'un serveur unique, comme décrit dans la figure 1.3.

De plus, l'interface des clients s'organise généralement en trois zones, présentées dans la figure 1.4 ; l'une exposant la conversation en cours, la seconde permettant la saisie d'un nouveau message et son envoi, et le troisième présentant les utilisateurs connectés (pour les interfaces de type IRC) ou bien la liste des contacts (pour les interfaces de type messagerie instantanée). Dans ce dernier cas, cette liste de contacts est souvent déportée sur une autre fenêtre ; pour des exemples, cf. figures 1.5 (liste de contacts) et 1.2 (messages et saisie).

2.1.4 Historique

On peut considérer la commande Unix *talk* comme le premier outil de tchat. Cette commande permettait à l'origine à plusieurs utilisateurs d'un même système, mais connectés sur des terminaux différents, de dialoguer en direct. En 1983, une nouvelle version permit de communiquer avec des systèmes distants [7], selon une architecture pair à pair. D'autres outils du même type, tels que *WinPopUp*, sur système Windows, ou encore Relay [Kel87], pour le réseau BitNet, virent le jour quelques années plus tard. Chaque utilisateur se voyait octroyer une fraction d'écran, et ses messages

[7] `http://en.wikipedia.org/wiki/Unix_talk`

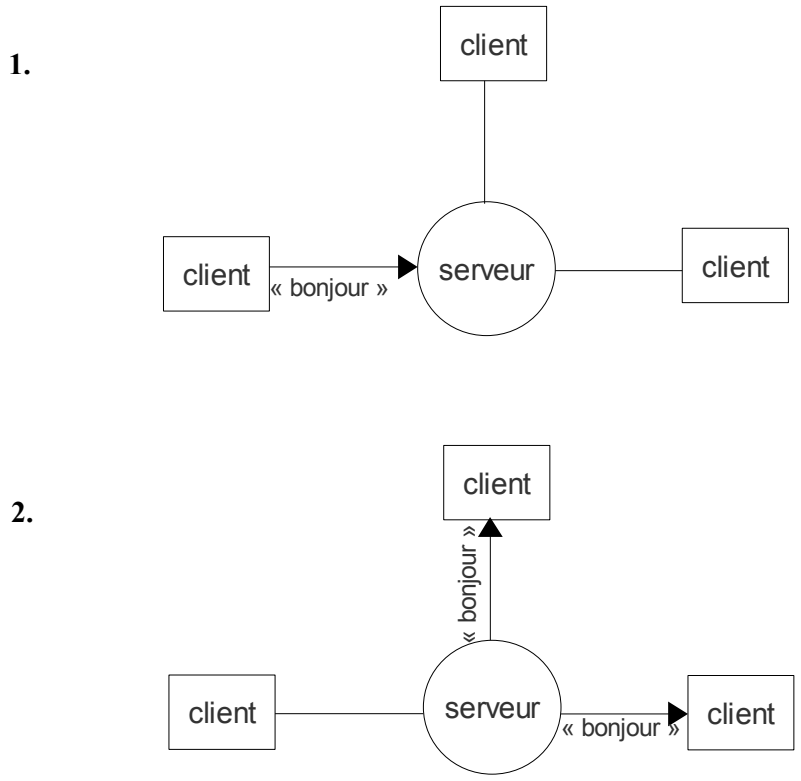


Figure 1.3 – Architecture client-serveur d'un système de tchat. Un serveur relaie les messages entre les clients, qui ne peuvent communiquer directement entre eux.

s'y ajoutaient séquentiellement, sans autre indication. De ce fait, il était impossible d'ordonner *a posteriori* les messages. Autre particularité, les messages étaient envoyés en temps réel au fur et à mesure qu'ils étaient tapés, et non après composition comme c'est le cas sur les outils postérieurs (figure 1.6). Les utilisateurs étaient identifiés par une séquence du type *nom d'utilisateur@URL de la machine*.

Progressivement, le système IRC [CL98], adapté pour Internet en 1988 par Jarkko Oikarinen [Oik93], prit le pas sur *talk*. Ce système repose sur une architecture différente, client-serveur, et est nettement plus élaboré : les discussions se déroulent sur des « canaux » dédiés, et les messages y apparaissent ordonnés par date de réception, comme dans une « vraie » conversation écrite. Les participants sont identifiés par des pseudonymes locaux, choisis lors de leur connexion au canal. Ainsi, un même pseudonyme peut fort bien être récupéré par un autre utilisateur par la suite. Exploitant cette apparente limitation, les utilisateurs retouchent fréquemment leur pseudonyme en fonction de leur activité ou de leur état d'esprit (par exemple en accolant un suffixe tel que _mange, _dort, _travail, _grognon à leur pseudonyme habituel). Le protocole étant ouvert, de nombreux clients et serveurs virent le jour, le plus utilisé aujourd'hui étant mIRC (figure 1.4).

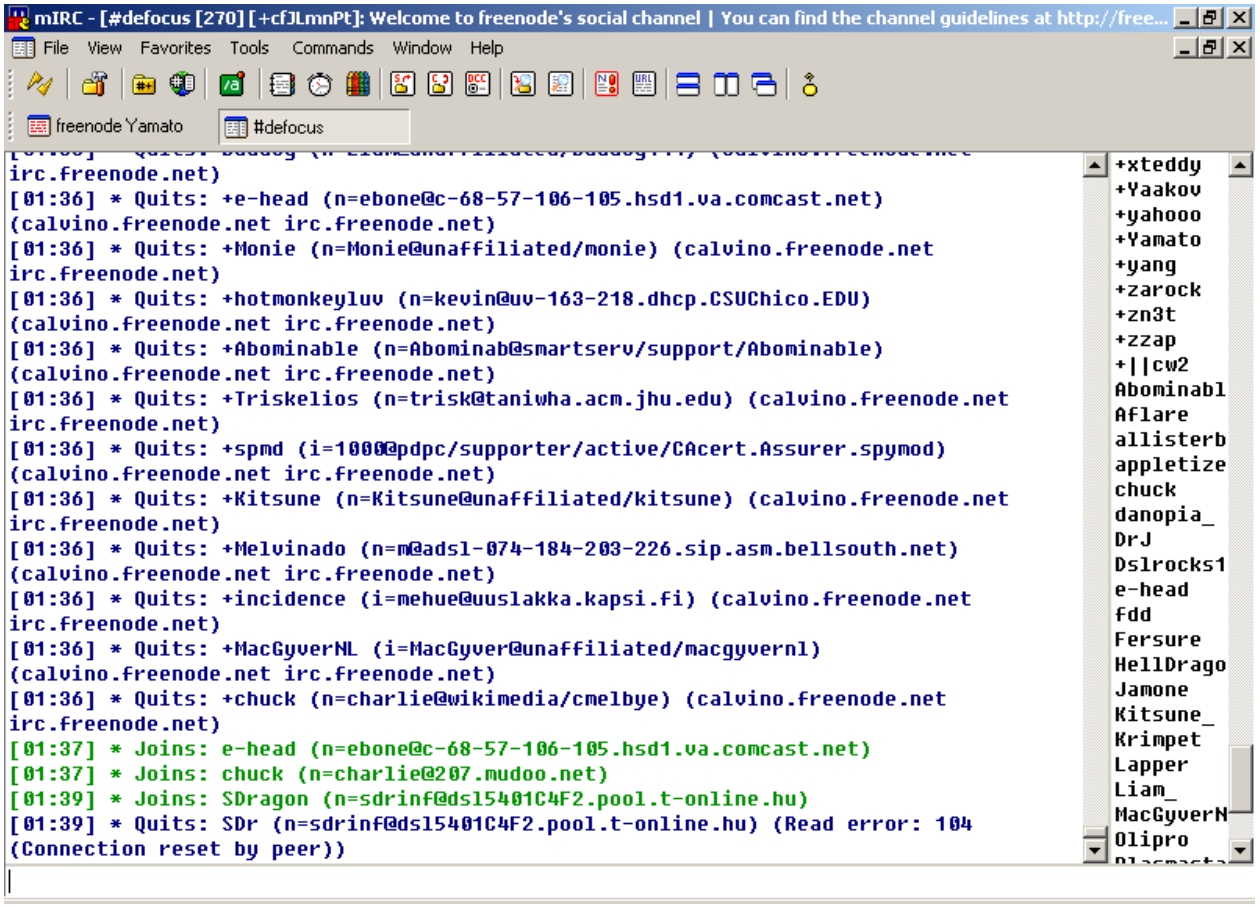


Figure 1.4 – Le client IRC mIRC, et l'interface classique en trois parties (source : Wikimedia Commons).

Guillaume Latzko-Toth a présenté un historique détaillé d'IRC dans deux articles [LT01a] [LT01b], mettant l'accent sur les aspects épistémologiques et ethnologiques de ce système. Ce protocole devint rapidement très populaire auprès des informaticiens, professionnels et étudiants [Oik93], qui enrichirent (et parfois pervertirent) le système en concevant des « robots » ou *bots*, des logiciels simples se faisant passer pour des utilisateurs normaux sur les canaux pour diverses raisons : « maintien en vie » du canal lorsque personne n'est connecté, contrôle des accès, statistiques, jeux, etc. ; mais aussi envoi de publicités, déversement de données aléatoires ou d'injures en boucle (*flood*), éjection voire déconnexion forcée d'utilisateurs ciblés, saturation de serveurs, etc. qui sont encore aujourd'hui très présentes dans l'image populaire que l'on se fait de l'informatique. C'est à partir de 1996 que les internautes américains « non-informaticiens » commencèrent à investir massivement les canaux IRC [LT01a], en faisant le premier système de tchat grand-public.

Toutefois, le règne d'IRC fut de courte durée, puisqu'en novembre 1996 Mirabilis, une société israélienne, publia ICQ[8] [icq], un logiciel qui associait pour la première

[8] Pour *I seek you*, « Je te cherche »

Figure 1.5 – Liste de contacts (source : Wikimedia Commons).

Figure 1.6 – La commande Unix *talk* (source : Wikimedia Commons)

fois un identifiant unique à chaque utilisateur, indépendamment de la machine utilisée par ce dernier : en l'occurrence un nombre donné par ordre croissant d'inscription au service, qui est octroyé une fois pour toutes. Rompant avec le principe d'un réseau organisé autour des salons de discussion propre à IRC, ICQ recentra l'utilisation sur

l'utilisateur, comme c'était le cas avec *talk*, mais en conservant l'architecture client-serveur et une interface proche d'IRC. L'outil central devint alors la liste de contacts, rendue possible par la permanence des identifiants : ce sont des gens que l'on connaît, dont le statut, connecté ou non, disponible ou non, est visible, et avec qui l'on peut initier des conversations dans une interface à la IRC (les conversations à plus de 2 utilisateurs sont possibles). En 1998, AOL[9], l'un des principaux fournisseurs d'accès à Internet américains, acheta Mirabilis et intégra à son offre un dérivé d'ICQ sous le nom d'AOL Instant Messenger (AIM), en faisant un outil très populaire aux États-Unis, où AIM est encore aujourd'hui le réseau de tchat le plus utilisé.

La même année, le protocole ouvert Jabber vit le jour. Il fut normalisé en 2004 par l'IETF[10], sous le nom XMPP. Créé par l'américain Jeremie Miller, il s'inspirait ouvertement d'ICQ [Mil99], dont il différait cependant par des identifiants alphanumériques à la *talk*, sur le modèle *identifiant@serveur*. Jusqu'à son adoption par Google pour son logiciel de tchat Google Talk, il resta surtout cantonné au monde de l'entreprise, soucieux de la confidentialité des données.

Aux côtés d'IRC, qui conservait toujours un noyau de fidèles, et des protocoles ICQ et Jabber, les années 1998-2000 ont vu apparaître et se développer une profusion de systèmes propriétaires, très semblables par leur architecture, leurs fonctionnalités de base et leur interface, mais aux protocoles incompatibles. Citons entre autres Yahoo ! Messenger en 1998, QQ (un clone d'ICQ) et MSN Messenger en 1999, et Gadu-Gadu en 2000. La Bulle Internet battait alors son plein, et l'objectif pour tous ces éditeurs était avant tout d'accumuler un maximum d'utilisateurs captifs (qui ne pouvaient pas communiquer avec les utilisateurs d'autres réseaux), le modèle économique étant basé sur l'envoi de publicités et l'offre de services annexes payants. La concurrence faisait rage pour capter un maximum d'utilisateurs et les nouveautés étaient nombreuses : « binettes », transfert de fichiers, multimodalité (dessin, voix, vidéo), diaporamas, partage d'agenda, de musique, etc. ; mais sans jamais bouleverser l'architecture des réseaux ou l'ergonomie des logiciels clients. Finalement en 2001, après la Bulle, le marché du tchat se trouva très fragmenté, situation que ne connaissaient pas le courriel et le Web par exemple [Min07]. Cette fragmentation se caractérisait par des situations d'hégémonie d'un protocole donné dans une région ou un pays donné. Face à ce problème, dans les années 2000, les clients multiprotocole, tels que Adium, GAIM/Pidgin, Kopete, Miranda, Proteus ou encore Trillian, ainsi que les passerelles, notamment de et vers le protocole Jabber, se multiplièrent. Mais la plupart des éditeurs de protocoles fermés, craignant de voir s'évanouir leurs utilisateurs, s'employèrent alors à bloquer ces logiciels et passerelles, et il fallut attendre 2004 pour que cessent à peu près définitivement ces tentatives de blocage.

En 2005, lorsque Google lance son logiciel Google Talk, celui-ci est basé sur (et compatible avec) le protocole ouvert Jabber/XMPP. Enfin, en septembre 2008, Jabber est racheté par Cisco, preuve d'un intérêt toujours aussi vif pour les technologies

[9] America On Line
[10] Internet Engineering Task Force

de tchat.

2.1.5 Situation actuelle

Aujourd'hui, la situation reste très fragmentée, avec de nombreux réseaux concurrents, qui sont toutefois utilisables de concert grâce aux clients multiprotocoles et aux passerelles. Le vénérable IRC est de plus en plus en plus délaissé par le grand-public, mais reste néanmoins toujours actif, et pas seulement auprès des informaticiens.

Wikipedia a rassemblé les données disponibles sur le nombre d'utilisateurs de chaque système. Toutefois, il s'agit soit d'estimations, soit de chiffres fournis par les éditeurs qui ont tendance à exagérer le nombre de leurs utilisateurs.

Protocole	Comptes	Utilisateurs	Année
Tencent QQ	783	318	2008
Winodws Live Messenger		294	2007
Yahoo ! Messenger		248	2008
Skype[11]	309		2008
AOL Instant Messenger	53	100	2006
Jabber	50		2007
eBuddy	35		2006
IBM Lotus Sametime	17		2007
ICQ	15		2006
Xfire	12		2008
MXit	7		2007
Gadu-Gadu		6	2008

Table 1.10 – Nombre d'utilisateurs de différents systèmes de tchat, en millions (estimations).

La société EQO Communications a réalisé en 2008 une étude sur la fragmentation des réseaux de tchat, dont le résultat est schématisé par la planisphère de la figure 1.7 [LaP08]. Cette étude montre que l'étude de Madeleine Pastinelli sur les canaux IRC [Pas99] est toujours d'actualité 9 ans plus tard, avec une importante fragmentation géographique des groupes d'utilisateurs ; un protocole majoritaire dans un pays donné peut être quasiment inexistant dans un autre. Ainsi MSN domine dans la plupart des pays, avec plus de 50% de part de marché en Europe, et plus de 60% en Amérique, avec deux exceptions notables sur ces continents : l'Allemagne qui est partagée entre ICQ (45%) et MSN (30%), la Russie où le rapport en faveur d'ICQ est encore plus fort qu'en Allemagne (ICQ 56% et YM 20%) et les États-Unis où le marché est divisé entre AIM (35%), YM (29%) et MSN (28%).

Certains systèmes sont spécifiques à un pays donné : ainsi, en Chine, Tencent QQ est utilisé par 60% des tchateurs, alors qu'il n'atteint pas 4% dans les autres pays. En dehors de cette étude, on peut aussi relever Gadu-Gadu, principalement utilisé

Figure 1.7 – Fragmentation du marché du tchat dans le monde (source : [LaP08], EQO Commmunications)

en Pologne (41% en 2008 [Mal08]) , MXit en Afrique du Sud (près de 5 millions d'utilisateurs [12]), Mail.ru en Russie (1 million - 30% [13]), et Psyc au Brésil (1 million d'utilisateurs [14]).

D'un point de vue académique, longtemps considérée comme « triviale », la langue du tchat ne fit guère l'objet d'études avant ces dernières années [LT01a], les chercheurs se concentrant dans un premier temps sur les aspects sociologiques [LT00].

2.1.6 Aspects sociaux

D'un point de vue social, la distinction entre tchat de type IRC et messagerie instantanée est pertinente. En effet, dans le premier cas, les utilisateurs se regroupent sur un canal thématique donné, et sont donc mieux à même de dialoguer avec des inconnus intéressés par la même thématique. Toutefois, contrairement à ce qu'on pourrait croire, des critères communautaires et géographiques subsistent et les canaux purement thématiques, c'est à dire n'incluant pas une grande part de discussions communautaires sans aucun rapport avec le thème officiel, sont rares. Les usagers d'IRC se connaissent souvent en dehors du tchat, et les canaux regroupent donc souvent des habitants d'une même région, voire d'une même ville [Pas99]. De ce point de vue, le tchat de type messagerie instantanée n'est que l'aboutissement de cette tendance à l'entre-soi. La messagerie instantanée ne prétend pas faire se rencontrer des inconnus partageant les mêmes centres d'intérêts, mais simplement permettre à des personnes se connaissant déjà de dialoguer, et se veut donc un calque des réseaux

[12] *MXit reaches 4.8 million milestone*, *BizCommunity*, `http://www.bizcommunity.com/Article/196/78/17027.html`

[13] *У Mail.Ru Agent 1 млн пользователей в сутки*, *C news*, `http://www.cnews.ru/news/line/index.shtml?2006/09/14/211037`

[14] *About Psyc*, `http://about.psyc.eu/Index`

sociaux préexistants. On peut d'ailleurs penser que l'effacement progressif de l'IRC au profit de la messagerie instantanée, et donc la perte d'un outil important dans la construction des réseaux sociaux, a facilité l'émergence des sites Internet « sociaux » et « Web 2.0 », la fonction créant l'outil.

Toutefois, l'utilisation du tchat ne se limite pas au Web « social » ; cet outil est utilisé dans d'autres contextes, notamment en entreprise. Cette utilisation peut être officieuse, les employés utilisant les outils classiques, ou bien officialisé par l'entreprise qui peut déployer des outils de Messagerie Instantanée en Entreprise (ou EIM pour *Enterprise Instant Messaging*) tels que *Lotus Sametime* d'*IBM* (1998) ; *Exchange Instant Messaging* de *Microsoft*, intégré dans *Office Communications Server* en 2007 ; ou encore l'*Extensible Communications Platform* de *Jabber*. Le monde de la finance a ses propres outils, comme *Reuters Messaging* et *Bloomberg Messaging*. Des études sur l'utilisation du tchat en entreprise [GD07] tendent à montrer que cela accroit la productivité des employés, leur permettant de mieux gérer et hiérarchiser les interruptions.

2.1.7 Aspects linguistiques généraux

Les caractéristiques de la langue du tchat ont été présentées en détail, notamment par Isabelle Pierozak [Pie03b], Émilie Guimier de Neef et Jean Véronis [GdNV04] pour le français, David Crystal [Cry01] et Jon Stevenson [Ste] pour l'anglais, et Michael Beißwenger [Bei01] et Susanne Krause [Kra97] pour l'allemand ; on n'en exposera donc ici que les grands principes.

Le « français tchaté » se caractérise en particulier par une syntaxe proche de la langue parlée, et surtout par sa graphie originale. Loin d'être une limitation, le caractère écrit des conversations de tchat semble en effet en être l'un des principaux attraits [Her99] [LT01b]. En fait, dans la langue du tchat, la graphie d'un lexème semble plus relever de la fantaisie de son auteur que de la norme orthographique, selon un processus de création lexicale permanente qu'Isabelle Pierozak [Pie03b] qualifie de « ludogénèse », et suffisamment souple pour permettre à certains utilisateurs de développer leur propre « voix » graphique.

Comme le souligne Isabelle Pierozak [Pie03b], la syntaxe des énoncés de tchat tient beaucoup de l'oral. Entre autres phénomènes propres à la langue parlée, les thématisations y sont fréquentes, ainsi que les constructions du type *situation + thème*[15] *+ (rhème*[16]*)*. On observe aussi des corrections (cf. table 1.11 (5) et (10)). Enfin, les messages longs sont souvent scindés en une succession de messages courts,

[15] En sémantique, on appelle *thème* l'élément d'un énoncé qui est réputé connu par les participants à la communication. La tradition linguistique anglophone emploie le terme de *topic*, qui se rencontre aussi réemployé en français, parfois francisé en topique (Wikipedia).

[16] En grammaire, le *rhème*, parfois également appelé propos, est un élément nouveau introduit dans l'énoncé, généralement par un déterminant indéfini ; par opposition à un thème qui est un élément connu de l'énoncé, généralement introduit par un déterminant défini. La tradition linguistique anglophone emploie les termes de *focus* ou *comment* (Wikipedia).

suivant les frontières propositionnelles, ce qui n'est pas sans rappeler les groupes prosodiques de la parole.

(1)	Scr1	ptit question si il y pas le meme nombre d'element ds les 22 liste tu fais comment ?
(2)	Scr1	car faut pas oublier que y auras pas que 2 conecction
(3)	Scr2	Scr1 : tu dois tjs avoir le meme nombre d'element :)
(4)	Scr1	faut
(5)	Scr1	faux
(6)	Scr1	enfin sauf lors d'une deco ou d'un connect
(7)	Scr2	mais a par pdt ces brefs instants a priori tu as pas de prob
(8)	Scr2	et puis au pire tu mets NULL dans le deuxieme membre de ta pere
(9)	Scr1	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)	Scr2	s/pere/paire/
(11)	Scr1	hum il est bizarre ton rezo :)

Table 1.11 – Exemple anonymisé de conversation entre Scr1 et Scr2 , sur le canal IRC #C++ [Fal05].

(1)	Scr1	Scr2 préfère sauce tarat
(2)	Scr1	tartare*
(3)	Scr2	du tou
(4)	Scr1	sisi
(5)	Scr1	a point
(6)	Scr3	poivre rulez
(7)	Scr4	sauce bernaise c la meilleure :)
(8)	Scr1	frite moelleuse
(9)	Scr3	nan a point sauce poivre
(10)	Scr3	ça rox
(11)	Scr1	lol
(12)	Scr3	moelleuse mais croustillante stp

Table 1.12 – Exemple anonymisé de conversation entre 4 scripteurs, sur le canal IRC #18-25ans [Fal05].

En ce qui concerne les caractéristiques propres du tchat, elles sont bien connues. Nous prendrons à titre d'exemple un extrait de dialogue sur un canal IRC de développeurs C++ (table 1.11).

Émoticons (ex. « :(», « :) » (3) (11)) : emblématiques de la communication électronique, les émoticons (aussi appelées *smileys*, frimousses, etc.).

Abréviations : beaucoup ne sont pas spécifiques à la communication électronique (ex. « tjs » (3) , « pdt » (7)), mais certaines le sont (« lol », « mdr »).

Graphie phonétique : l'orthographe suit la prononciation (« salut les zamis », « ptit » (1) , « rezo » (11)).

Graphie atténuée : on relève par ailleurs fréquemment une graphie que l'on pourrait qualifier d'atténuée, transcrivant des variations phonologiques (« kikoo » pour « coucou », « oki » pour « okay », etc.). Il s'agit d'une modification de la graphie d'un mot, destinée à lui donner une prononciation légèrement différente de la norme. Parfois, ces variantes phonologiques se doublent d'allongements vocaliques, comme dans « kikooooooooo » par exemple.

Les logiciels de messagerie récents enrichissent ces divers effets à l'aide d'icônes, de polices de caractères, couleurs, etc. mais cela reste fondamentalement la même chose.

Ces « effets », volontaires, visent plusieurs objectifs :

Spontanéité : il s'agit de briser le caractère formel de la conversation écrite classique (courriel, courrier), afin de paraître plus spontané, plus sincère ; ce qui revient à instituer un autre formalisme. Nous proposons de rapprocher ce phénomène de celui consistant à feindre de se montrer outrageusement grossier avec des proches afin de marquer sa spontanéité et sa sincérité, décrit par Catherine Kerbrat-Orecchioni [KO05] pour les conversations entre « jeunes ».

Atténuation : des messages écrits peuvent facilement paraître secs, presque violents. Il s'avère donc nécessaire de les « affaiblir » à l'aide des divers effets décrits *supra*, mais aussi de leur donner un caractère oral, informel (rejoignant l'objectif de spontanéité). L'effet recherché semble similaire à celui des *adoucisseurs* de la langue parlée (par exemple dans « une petite pièce », cf. [KO05]).

Expressivité : par défaut, un texte dactylographié est parfaitement anonyme. Les tchateurs ressentent le besoin d'exprimer leur attitude, voire de développer une « voix » graphique qui les distingue des autres. Cela est bien entendu à rapprocher de l'importance de l'expressivité dans le dialogue, que nous avons rappelée précédemment.

Rapidité : les conversations de tchat se veulent en temps réel, comme dans un dialogue oral, où le temps de la planification déborde sur le temps de la production. Il faut réduire autant que possible le délai entre chaque message. Il est toutefois important de ne pas confondre cette dernière motivation avec la notion proche de concision, pertinente pour les SMS mais pas dans le cas du tchat ; contrairement aux SMS, les messages de tchat ne sont pas facturés en fonction de leur taille et sont entrés à l'aide d'un clavier bien plus ergonomique. Il s'agit donc de composer rapidement les messages, avec peu de planification préalable, mais pas de passer du temps à composer des messages synthétiques.

Il faut souligner néanmoins que malgré toutes ces divergences par rapport à l'écrit « standard », la finalité du tchat demeure le dialogue, et les tchateurs sont soucieux, jusqu'à un certain point, de la clarté de leur messages. En témoignent les corrections effectuées a posteriori par les tchateurs eux-même, généralement lorsqu'un mot mal orthographié peut être confondu avec l'un de ses homophones hétérographes (« s/pere/paire/ »). Nos observations sur corpus tendent à montrer que

plus le dialogue est finalisé et plus les locuteurs sont actifs dans la durée (et donc expérimentés), plus les productions sont lisibles, à défaut de totalement rigoureuses sur le plan formel.

2.1.8 Corpus disponibles

Du fait de cet intérêt encore récent, peu de corpus disponibles atteignent une taille critique [BS07], tant du point de vue quantitatif (nombre de formes) que qualitatif (nombre de canaux). Les corpus sont souvent petits, ne servent que pour un projet donné, et ne sont pas diffusés. Étant donné le caractère privé des conversations de type messagerie instantanée, il est difficile d'en collecter des corpus ; tous les corpus disponibles proviennent donc de discussion IRC qui, elles, sont publiques. Certains de ces corpus sont dans un simple format textuel, d'autres sont structurés en XML avec des annotations *techniques* (date, heure, auteur, message automatique ou produit par un véritable utilisateur, etc.), parfois (rarement) *syntaxiques* et *pragmatiques*.

Il faut tout d'abord citer les travaux précurseurs de la société Eulogos [17], qui a collecté en 1997-1998 un corpus de près de près de 75 000 messages pour l'italien, librement consultable sur Internet. En 2004, nous avons collecté ce qui reste le principal corpus de tchat français, structuré, avec plus de 4 millions de messages, sur plus de 100 canaux différents [Fal05] (cf. *infra*). En 2006, l'université de Dortmund a publié un corpus de 150 000 messages en allemand [18], et l'université de Tartu un corpus pour l'estonien, intégré dans l'*Estonian Reference Corpus*. Enfin, on notera l'absence de grand corpus de tchat pour l'anglais, avec néanmoins une initiative très intéressante datant de 2008 : le *Naval Postgraduate School Chat Corpus*, qui ne contient actuellement qu'un peu plus de 10 000 messages (mais d'après les auteurs ce nombre devrait augmenter), mais avec des annotations manuelles d'ordre morphologique (catégories lexicales) et pragmatique (actes de langage) [FM07].

Langue	Messages	Formes	Canaux	Format	Accès
italien	74 758	847 223	14	texte	libre
français	4 192 033	23 011 876	105	XML, technique	libre [1]
allemand	150 060	1 150 001		XML, technique	libre [2]
estonien	ca. 2 800 000	ca. 7 000 000		XML, technique	libre
anglais	10 567		1	XML, technique, syntaxique, pragmatique	libre

Table 1.13 – Nombre d'utilisateurs de différents système de tchat, en millions (estimations).

[17] *Corpus di conversazioni da chat-line in lingua italiana, da registrazioni effettuate nel primo trimestre 1998*, `http://www.intratext.com/X/ITA0192.HTM`, 2001

[18] *Dortmunder Chat-Korpus* : `http://www.chatkorpus.uni-dortmund.de`

[1] Accès partiel : 2 523 321 messages.

[2] Accès partiel : 59 876 messages.

On signalera aussi, pour un autre type de texte souvent assimilé (à tort, comme nous le verrons) au tchat, le corpus de 30 000 SMS en français « SMS pour la science » collecté à l'université de Louvain [FKP06].

2.2 Étude d'après corpus

2.2.1 Collecte d'un corpus de tchat

La collecte d'un corpus de tchat ne va pas sans soulever certaines questions éthiques. Il s'agit en effet d'enregistrer des conversations qui, si elles sont publiques, n'en sont pas moins en principe éphémères. De plus, à moins de créer un canal dédié à la collecte d'un corpus, il n'est pas possible de prévenir tous les utilisateurs du fait qu'ils vont être enregistrés. Nous avons donc choisi de constituer ce corpus à partir de logs (ou « archives ») librement consultables sur Internet, plutôt que d'enregistrer nous-même des sessions de tchat. Un autre avantage de cette méthode est qu'elle permet de récupérer en une seule fois une grande quantité de données.

Le corpus de tchat est constitué à partir des logs disponibles sur le site *bots-tats.com*. Ce site publie les logs de quelques centaines de canaux du serveur IRC EpikNet. La publication des logs est un choix de la part du créateur du canal ; et ce dernier peut en outre restreindre leur consultation aux utilisateurs enregistrés. Toutefois, quelques créateurs de canaux (une centaine) ont décidé de les rendre accessibles à tous, et ce sont les logs de ces canaux que nous avons regroupés au sein du corpus. Les logs, consultables sous forme de pages HTML sur Internet (cf. figure 1.8), sont tout d'abord extraits à l'aide d'*Httrack*, un « aspirateur de sites ». On obtient des fichiers HTML classés par canal, comportant environ 1 000 messages. À l'intérieur d'un fichier HTML, les messages sont classés par date et comportent la date et l'heure de réception (à la minute près, de nombreux messages consécutifs peuvent donc avoir la même heure), mais les fichiers eux-mêmes ne sont pas classés. De plus, certains fichiers se chevauchent, un même message peut donc être présent plusieurs fois. De plus, les messages des utilisateurs sont noyés au milieu de notifications du système (connexions d'utilisateurs, messages automatiques, etc.), et une classification des messages est donc un préalable avant toute analyse linguistique. On effectue les traitements suivants :

1. Les messages et les informations relatives (pseudonyme, date, heure) sont extraits à l'aide d'expressions régulières.
2. Les messages sont regroupés par canal.
3. Les messages sont typés automatiquement en tant qu'événement (pas d'auteur mentionné), commande (commençant par le caractère « ! »), ou à défaut message « normal ».
4. En ce qui concerne les événements, aisément discriminables à l'aide d'expressions régulières, un second typage est effectué : connexion, déconnexion, changement de pseudonyme, etc. ; ce typage s'accompagne d'informations spécifiques, par exemple pour le changement de pseudonyme, pseudonymes source et cible.

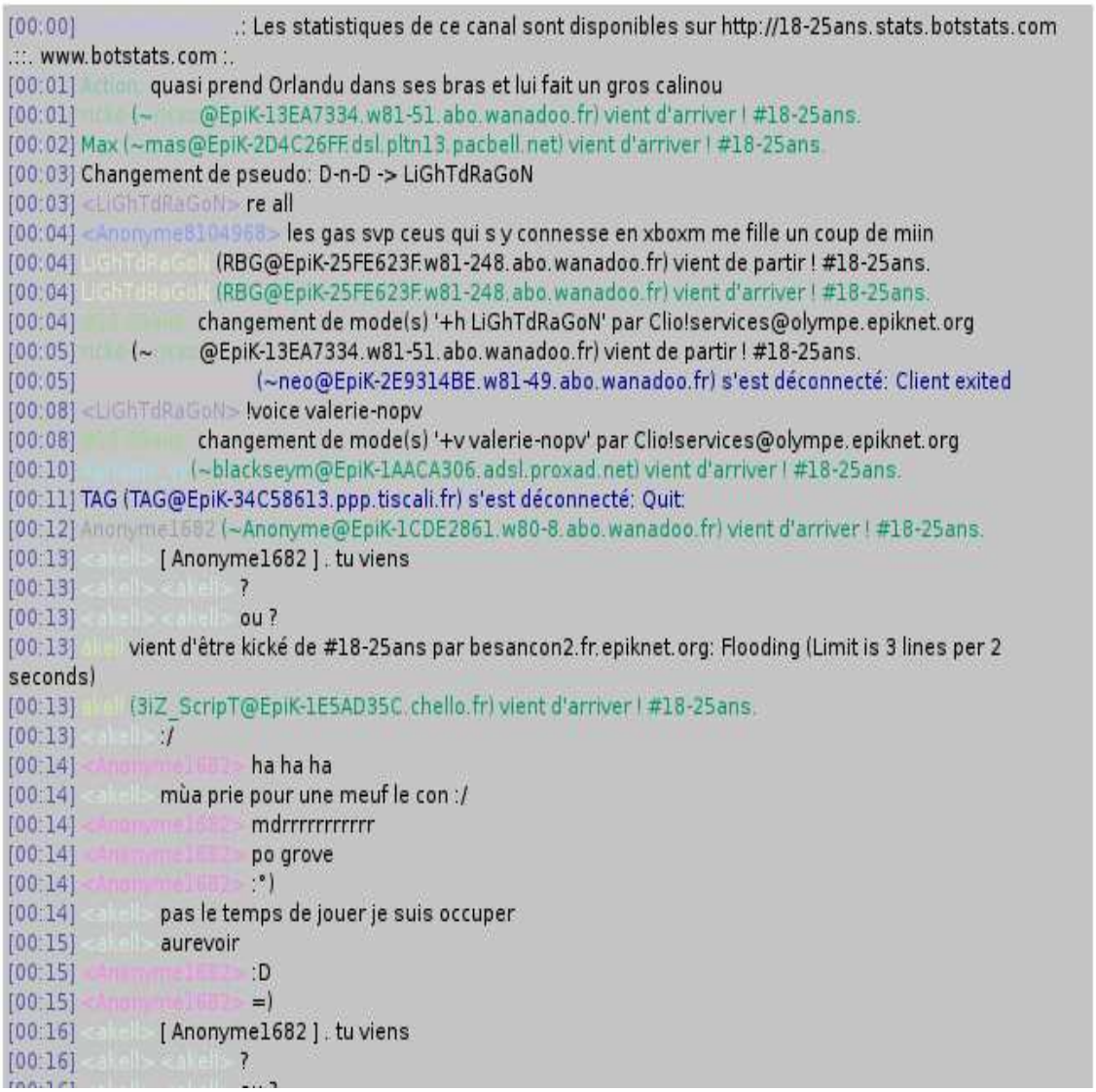


Figure 1.8 – Un aperçu des archives du canal #18-25ans telles qu'elles apparaissent dans *botstats.com*

5. On l'a vu, les utilisateurs changent fréquemment de pseudonyme au cours d'une session, et après déconnexion n'importe qui peut se réapproprier le pseudonyme. Afin de pouvoir suivre les scripteurs, un identifiant unique basé sur leur identifiant de connexion, signalée par un message de type événement, est calculé et leur est attribué.

6. Des statistiques sur le nombre d'occurrences des différentes formes et le nombre d'interventions des scripteurs sont générées, dans un format simple de type « *clef*=*valeur* ».

7. Enfin, un fichier « canal » au format XML est généré, ainsi qu'un fichier de requêtes SQL pour l'import du corpus dans une base de données MySQL.

L'enchaînement précis de ces traitements est décrit dans la figure 1.9.

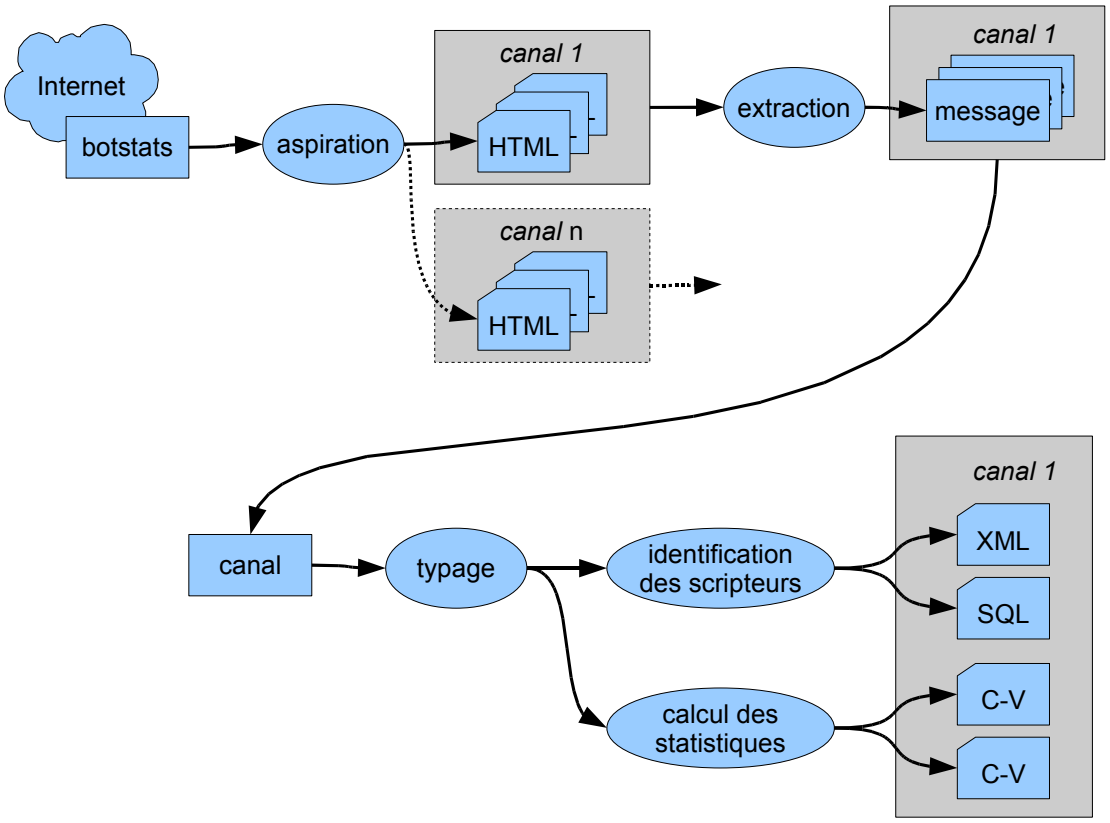


Figure 1.9 – Constitution du corpus.

2.2.2 Format du corpus

L'activité d'un canal de tchat peut être représentée par une succession de messages, produits par différents auteurs. Outre les messages « normaux », rédigés par un auteur humain à destination de lecteurs humains, il faut distinguer quelques cas particuliers :

les commandes, qui sont destinées au serveur (afficher la liste des utilisateurs par exemple) ou à un robot (fréquent dans le cadre de jeux textuels comme le « pendu »), et qui appartiennent à un langage formel ;

les événements, déclenchés à l'aide de raccourcis clavier ou lors de certains événements (déconnexion de l'utilisateur par exemple).

Notre format de codage XML rend compte de cette structure, et inclut diverses indications techniques : date, heure, auteur. Il est détaillé dans la figure 1.10.

Notre corpus est consultable en ligne[19]. Une interface PHP/HTML permet de sélectionner des fragments de corpus et de reconstituer leur format XML, en fonction des critères entrés par l'utilisateur *via* un formulaire. Afin d'assurer des performances raisonnables, nous avons dû intégrer notre corpus XML dans une base de données

[19] www-clips.imag.fr/geta/User/achille.falaise

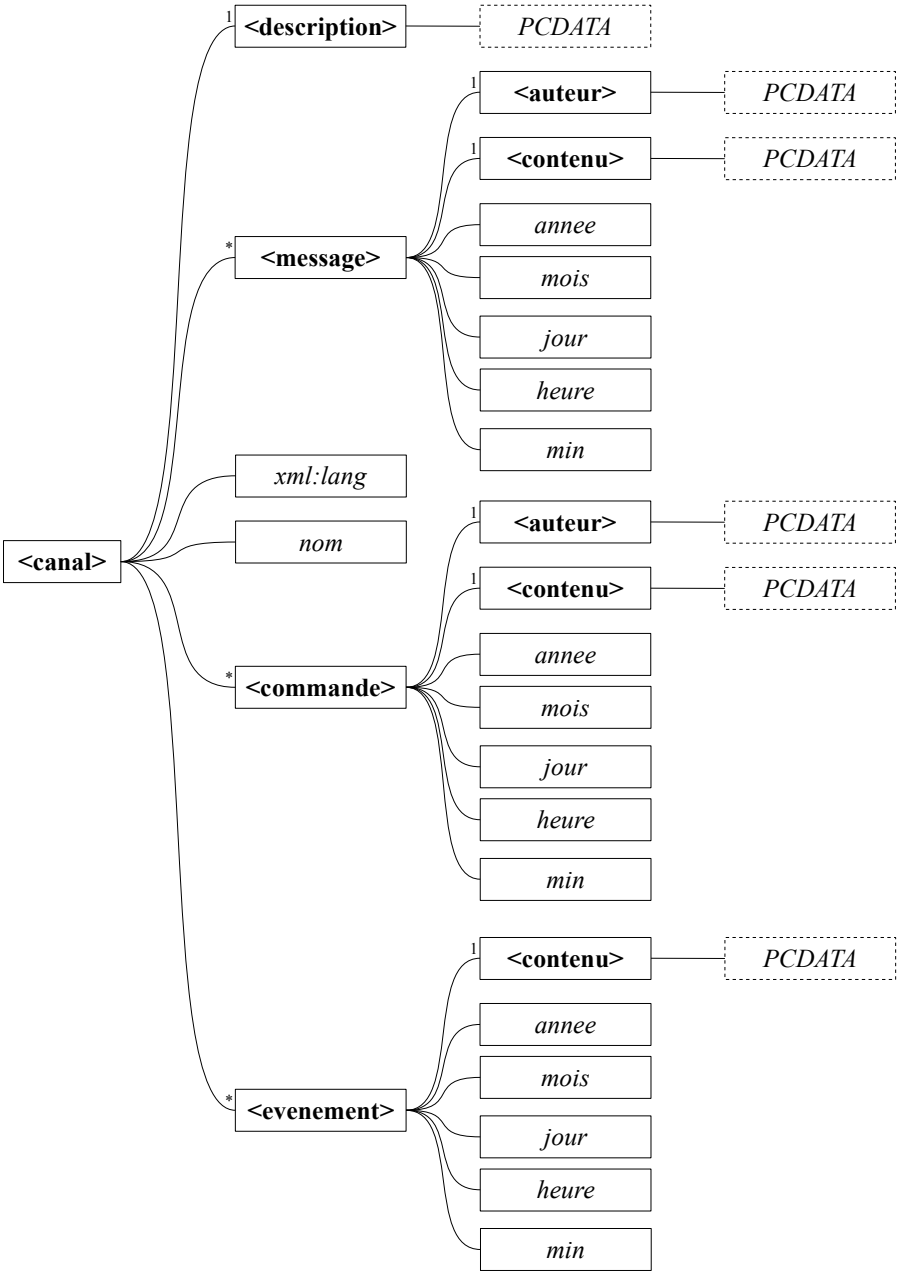


Figure 1.10 – Format de chaque canal du corpus.

MySQL. L'affichage est assuré par une feuille de style XSL, qui convertit le code XML en HTML (cf. figure 1.11). Un menu permet de sélectionner un canal en particulier, ou bien tous les canaux. On peut ensuite choisir le nombre d'interventions par page ainsi que leur type. Les interventions sont classées par date, et colorées en fonction de leur type. Pour chaque intervention, l'identifiant du scripteur ainsi

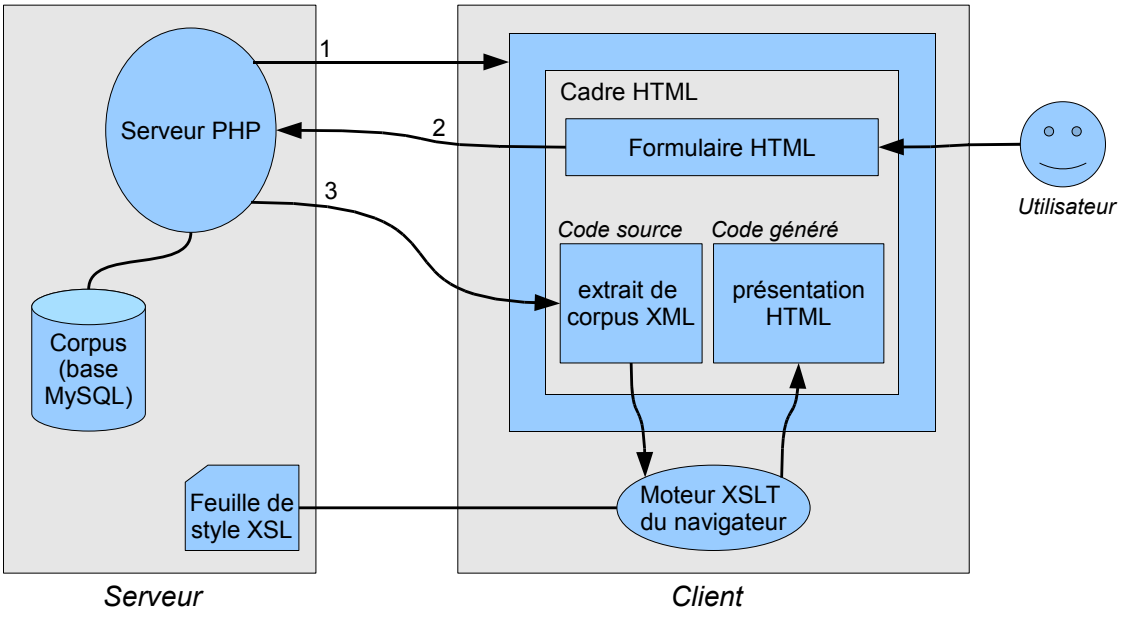
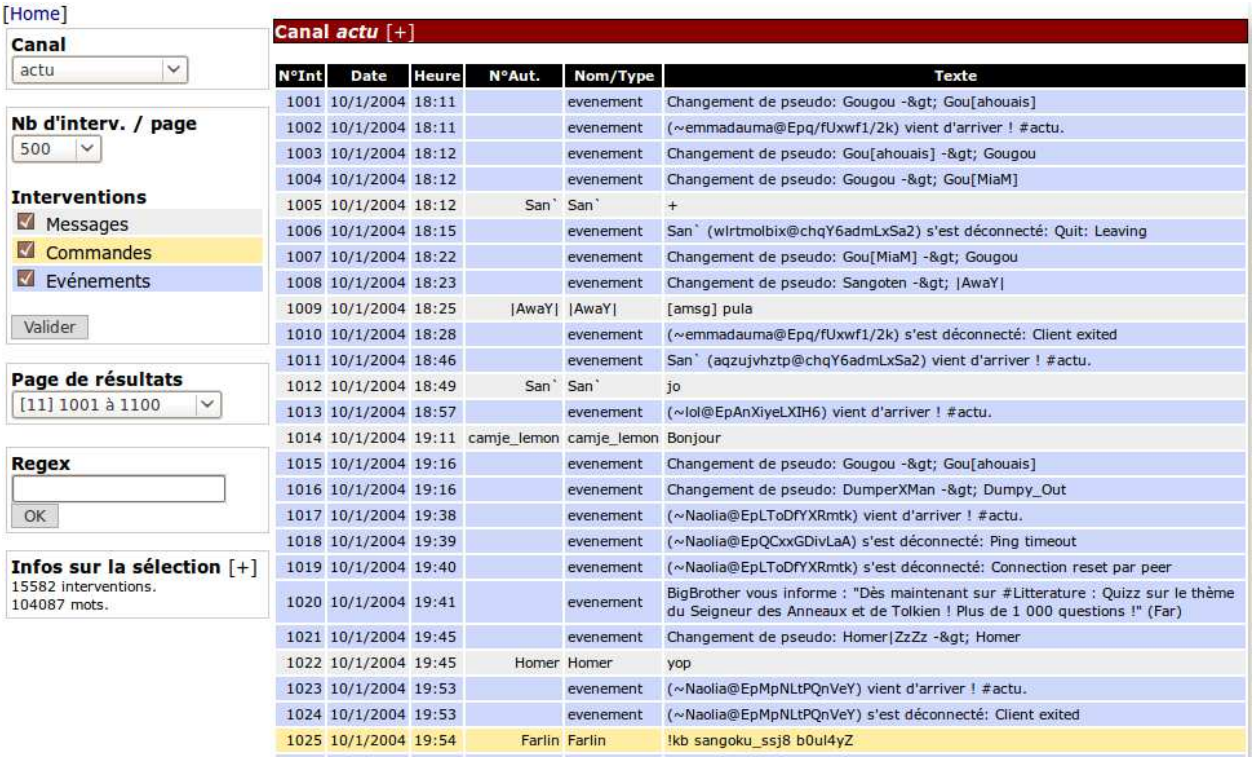


Figure 1.11 – Principe de fonctionnement du site de consultation du corpus.



[Home]

Canal
actu

Nb d'interv. / page
500

Interventions
- Messages
- Commandes
- Evénements

Valider

Page de résultats
[11] 1001 à 1100

Regex
OK

Infos sur la sélection [+]
15582 interventions.
104087 mots.

Canal *actu* [+]

N°Int	Date	Heure	N°Aut.	Nom/Type	Texte
1001	10/1/2004	18:11		evenement	Changement de pseudo: Gougou -> Gou[ahouais]
1002	10/1/2004	18:11		evenement	(~emmadauma@Epq/fUxwf1/2k) vient d'arriver ! #actu.
1003	10/1/2004	18:12		evenement	Changement de pseudo: Gou[ahouais] -> Gougou
1004	10/1/2004	18:12		evenement	Changement de pseudo: Gougou -> Gou[MiaM]
1005	10/1/2004	18:12	San`	San`	+
1006	10/1/2004	18:15		evenement	San` (wlrtmolbix@chqY6admLxSa2) s'est déconnecté: Quit: Leaving
1007	10/1/2004	18:22		evenement	Changement de pseudo: Gou[MiaM] -> Gougou
1008	10/1/2004	18:23		evenement	Changement de pseudo: Sangoten -> |AwaY|
1009	10/1/2004	18:25	|AwaY|	|AwaY|	[amsg] pula
1010	10/1/2004	18:28		evenement	(~emmadauma@Epq/fUxwf1/2k) s'est déconnecté: Client exited
1011	10/1/2004	18:46		evenement	San` (aqzujvhztp@chqY6admLxSa2) vient d'arriver ! #actu.
1012	10/1/2004	18:49	San`	San`	jo
1013	10/1/2004	18:57		evenement	(~lol@EpAnXiyeLXIH6) vient d'arriver ! #actu.
1014	10/1/2004	19:11	camje_lemon	camje_lemon	Bonjour
1015	10/1/2004	19:16		evenement	Changement de pseudo: Gougou -> Gou[ahouais]
1016	10/1/2004	19:16		evenement	Changement de pseudo: DumperXMan -> Dumpy_Out
1017	10/1/2004	19:38		evenement	(~Naolia@EpLToDfYXRmtk) vient d'arriver ! #actu.
1018	10/1/2004	19:39		evenement	(~Naolia@EpQCxxGDivLaA) s'est déconnecté: Ping timeout
1019	10/1/2004	19:40		evenement	(~Naolia@EpLToDfYXRmtk) s'est déconnecté: Connection reset par peer
1020	10/1/2004	19:41		evenement	BigBrother vous informe : "Dès maintenant sur #Litterature : Quizz sur le thème du Seigneur des Anneaux et de Tolkien ! Plus de 1 000 questions !" (Far)
1021	10/1/2004	19:45		evenement	Changement de pseudo: Homer|ZzZz -> Homer
1022	10/1/2004	19:45	Homer	Homer	yop
1023	10/1/2004	19:53		evenement	(~Naolia@EpMpNLtPQnVeY) vient d'arriver ! #actu.
1024	10/1/2004	19:53		evenement	(~Naolia@EpMpNLtPQnVeY) s'est déconnecté: Client exited
1025	10/1/2004	19:54	Farlin	Farlin	!kb sangoku_ssj8 b0ul4yZ

Figure 1.12 – Interface de consultation du corpus : premiers messages du canal #actu.

que son pseudonyme sont donnés (cf. figure 1.12). Il est possible d'effectuer des filtrages par mot ou expression régulière (cf. figure 1.13). Le fait de cliquer sur une

N°Int	Date	Heure	N°Aut.	Nom/Type	Texte
1381	11/1/2004	1:44	Phara0n	Phara0n	Mais ya pas de chat
1387	11/1/2004	1:45	Womby	Womby	chat c'est un bouffe tune inutile gprs oui c'est un manque video bah perso deja je vois pas l'utilité de la photo
2007	13/1/2004	12:33	Womby	Womby	j'imagine que vous alliez rue de brabant pour vos achats donc
2186	13/1/2004	13:21	Womby	Womby	il obtient le rôle principal dans l'excellent Panic in needle park (1970) de Jerry Schatzberg, polar qui se déroule dans le milieu toxicomane à Manhattan
2193	13/1/2004	13:26	StatsBot	StatsBot	Si vous voulez utiliser un bot de stats simple et complet, visitez http://stats-bot.serveftp.org:81/ : Inscrivez vous et ajoutez votre chan. Merci et bon chat.
2980	17/1/2004	19:4	San`	San`	c un vrai chat
5102	25/1/2004	20:37		evenement	BigBrother vous informe : "Participez au concours de poèmes de la St-Valentin et envoyez nous votre plus beau poème d'amour sur stvalentin@epCq9zTPVREjU ! Un Pack EpiK et un Bon d'Achat de 30€ sur la Boutique EpiK (http://boutique.epiknet.org) à la clé !" (gohan)
5916	28/1/2004	13:38	enter	enter	qui veut tchaté avec moi
5958	28/1/2004	15:52	enter	enter	vous pouvez tchaté quand vous voulez avec moi
6008	28/1/2004	17:24	paige	paige	qui tchat avec moi stp
6032	28/1/2004	18:23	enter	enter	quelqu'un veut tchaté avec moi
6039	28/1/2004	18:25	enter	enter	qui veut tchaté avec moi
6050	28/1/2004	18:49	Anonyme5162146	Anonyme5162146	alors je voudrais tchaté les mecs
6062	28/1/2004	19:3	Womby-Aw	Womby-Aw	ben en gros je reste présent pour la discussion mais comme on est dans un salon de chat parlant d'actualité je vais allez lire les news (l'actualité)

Figure 1.13 – Interface de consultation du corpus : filtrage par expression régulière sur le canal #actu, pour afficher différentes graphies de « tchat ».

intervention permet alors d'en afficher le contexte. Enfin, l'interface étant générée à l'aide d'une feuille de style XSLT, le code XML correspondant à la sélection est toujours disponible dans le navigateur en affichant le code source.

2.2.3 Évaluation du corpus

D'un point de vue quantitatif, la somme de données collectées est assez considérable : 4 192 033 messages, répartis sur 105 canaux de tchat. Si l'on considère un mot comme une suite de caractères délimitée par les signes de ponctuations traditionnels (cette définition n'est pas forcément la plus adaptée au tchat, mais est acceptable en première approche), alors le corpus comporte 23 011 876 mots, soit une moyenne d'environ 5,5 mots par message. De ce point de vue, ce corpus apparaît sans commune mesure avec l'existant. A titre d'exemple, le plus important corpus auquel nous avons eu accès est le corpus de tchat italien constitué par la société Eulogos [Eul01], qui comporte 849 510 mots.

Constituer un corpus aussi important n'a de sens que si les données collectées sont suffisamment hétérogènes. De ce point de vue, le nombre important de canaux en jeu semble une bonne garantie : les thèmes abordés par les canaux sont variés, et vont du tchat généraliste où l'on discute de tout et de rien, au tchat spécialisé dans les problèmes de programmation, ou encore les débats concernant l'actualité. On relève aussi des différences d'ordre pragmatique. En plus des traditionnels bavardages, certains canaux sont plus ou moins dédiés aux jeux (pendu, quizzes), alors que dans d'autres la conversation est alimentée par des dépêches AFP. D'autres enfin sont consacrés à des discutions techniques, sous forme de question/réponse, par exemple

sur un canal consacré aux questions de programmation. Certains canaux semblent à première vue assez originaux sur le plan linguistique. Ainsi, on peut constater en comparant les figures 1.14 et 1.15 que le canal #edelweiss comporte peu de formes pour sa taille (42% de moins que #ffparadise), de taille pourtant plus réduite.

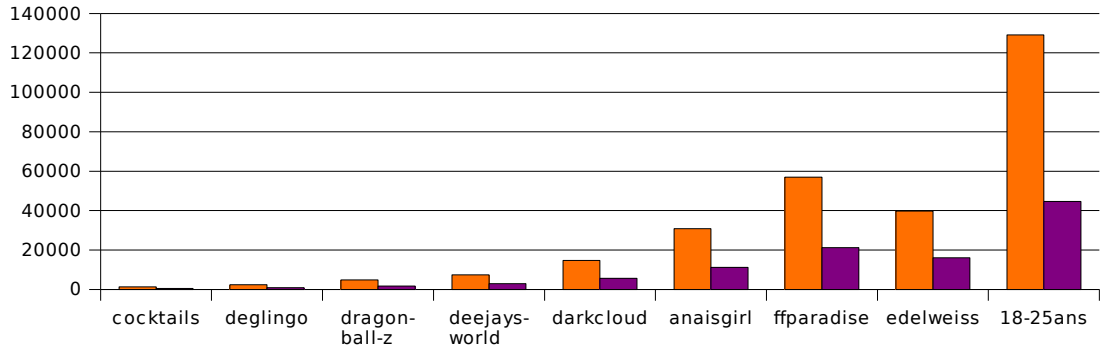


Figure 1.14 – Nombre de formes (barre de gauche), et nombre de formes présentes au moins deux fois (barre de droite), dans quelques canaux du corpus.

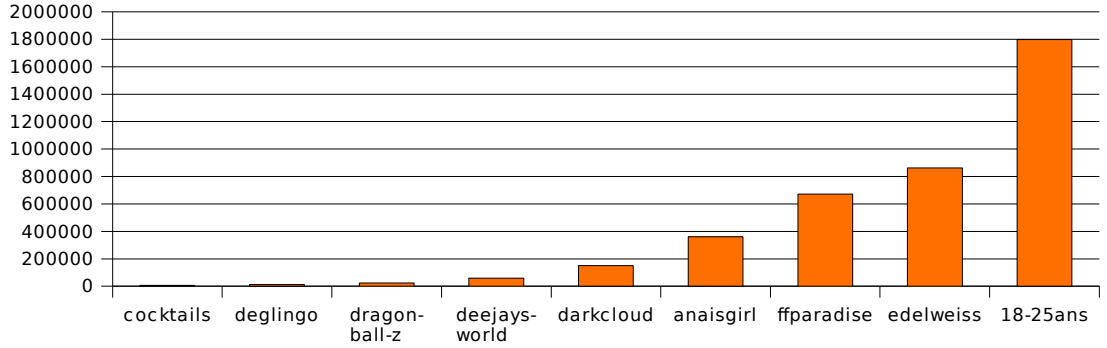


Figure 1.15 – Nombre de mots dans quelques canaux du corpus ; par « mot», on entend un ou plusieurs caractères entre deux blancs (la ponctuation, utilisée dans les émoticons, n'est pas prise en compte).

2.2.4 Tchat et SMS

On rapproche souvent tchat et SMS. On l'a déjà mentionné, sur le plan de l'ergonomie, ces outils sont assez différents : clavier ergonomique, messages de taille illimitée et gratuits, quasi-synchronicité pour le tchat ; petit clavier à 12 touches, messages limités à 160 caractères et payants, temps d'attente pour le SMS. Le corpus de tchat collecté, par comparaison au corpus du projet « SMS pour la science » [FKP06], nous permet de déterminer si cette différence au niveau des outils se retrouve au niveau des productions textuelles.

Nous avons comparé un extrait de 1 849 111 messages, issu du corpus de tchat, avec quelques variantes graphiques relevées sur un extrait de 30 000 messages du corpus de SMS [FKP06] (cf. table 1.14).

Graphie	SMS : absolu	SMS : relatif	Tchat : absolu	Tchat : relatif
pTr[1]	0	0	0	0
p e	14	46,67	3	0,16
p ê	0	0	0	0
p-e	277	923,33	116	6,27
pit etre	1	3,33	9	0,49
ptèt	11	36,67	2	0,11
ptète	0	0	1	0,05
pt-etre	2	6,67	0	0
pe	180	600,00	8	0,43
pe etre	0	0	56	3,03
pe tetre	0	0	8	0,43
ptetr	67	223,33	6	0,32
ptetre	40	133,33	143	7,73
ptètre	0	0	4	0,22
ptet	199	663,33	543	29,37
peu etre	18	60,00	5	0,27

Table 1.14 – Nombre d'occurrences de diverses graphies de « peut-être » en SMS et en tchat, relevés sur des extraits de 30 000 et 1 849 111 messages respectivement, en valeurs absolue et relative aux nombres de messages totaux, en occurrences pour 100 000.

Les messages sont généralement plus longs en SMS qu'en tchat, aussi le calcul du nombre d'occurrences doit être effectué avec précaution. On voit toutefois clairement que certaines variantes graphiques sont privilégiées par le tchat, tandis que d'autres le sont par le SMS.

Bien qu'on puisse effectivement poser comme point commun entre tchat et SMS la présence de variantes graphiques volontaires, on voit se dessiner un différence majeure dans la distribution de ces variantes. Ainsi, en SMS, la graphie « ptetr » est 25% plus fréquente que « ptetre », tandis qu'en tchat, c'est très nettement l'inverse : « ptetre » est 96% plus fréquente que « ptetr ». Le tchat va privilégier des variantes euphoniques[20], là où le SMS va préférer réduire le nombre de caractères.

2.2.5 Tchat et graphies peu normées

Si aujourd'hui on considère le respect strict d'une norme comme une caractéristique de la langue écrite, cela n'a pas toujours été le cas. D'un point de vue his-

[1] Forme donnée par un « dictionnaire du langage SMS ».

[20] Dans les domaines de la linguistique et de la phonétique, l'euphonie est une qualité des sons agréables à entendre ou aisés à prononcer. (...) L'*euphonie* relève de la fonction poétique dans le schéma de Jakobson (Wikipedia).

torique, la plupart des systèmes d'écriture sont relativement peu normés comparés aux système modernes, et évoluent *parfois* vers une normalisation. Comme l'a fait remarquer [GdNV04], on voit réapparaître en tchat des phénomènes de créativité phonético-graphique qui relèvent de l'écrit non normé.

En l'absence d'autorité linguistique centralisée, la graphie d'un même terme peut varier suivant la fantaisie du scripteur. Ainsi, en glyphes mayas, un même mot peut s'écrire de diverses manières [Rau02], au choix du scripteur (figure 1.16). C'est aussi le cas en chinois archaïque (figure 1.17), de même qu'avec les hiéroglyphes égyptiens (figure 1.18). Enfin, même des langues dont la graphie s'appuie plus sur la pho-

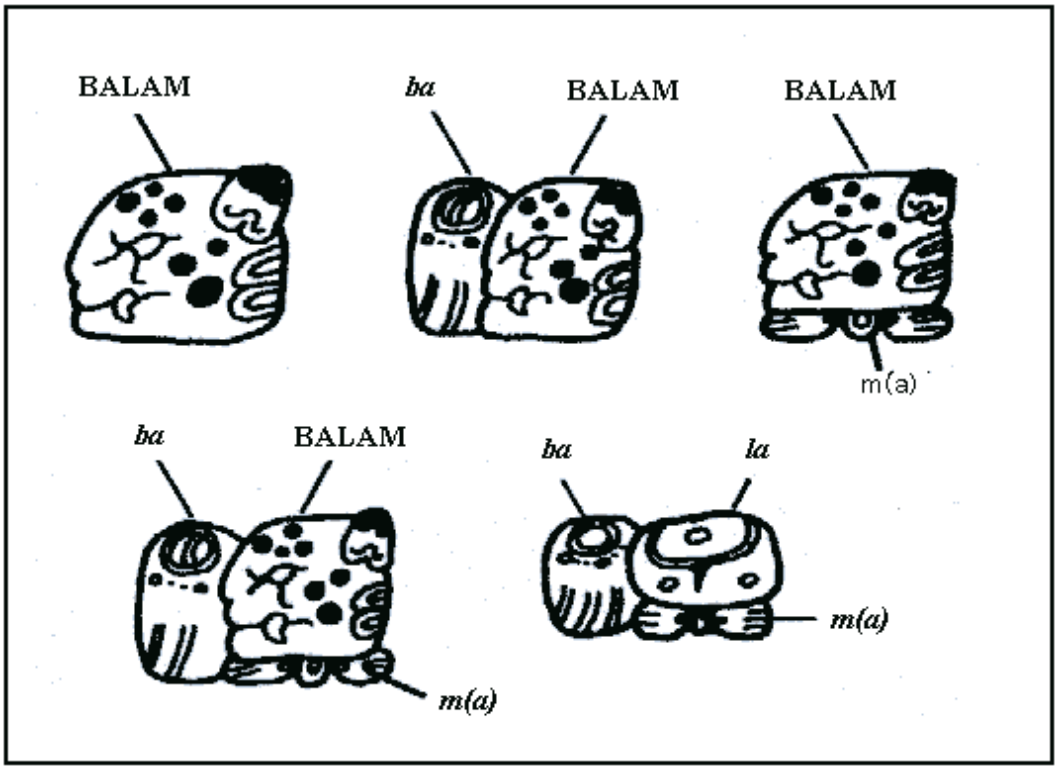


Figure 1.16 – Cinq manières d'écrire le mot *balam* (jaguar) en maya : idéographique (premier glyphe), idéo-syllabique (trois glyphes suivants), et syllabique (dernier glyphe) [Rau02].

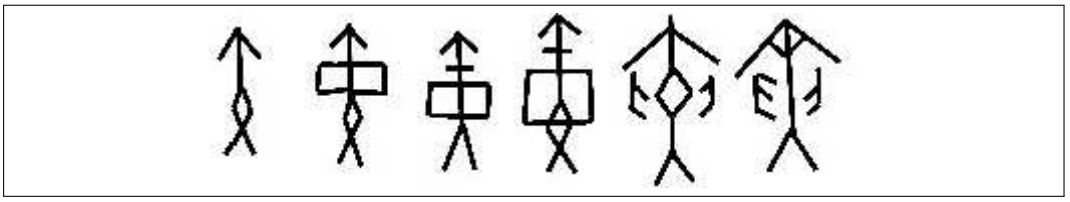


Figure 1.17 – Différentes façons d'écrire yín (troisième branche terrestre du cycle sexagésimal chinois, source : *Wikimedia Commons*).

Figure 1.18 – Différentes façons d'écrire *nfr* (« beau, bon, parfait »).

nétique, comme le français, peuvent faire l'objet de graphies variables, dès lors que les signes disponibles ne couvrent pas tout le spectre phonologique. Ainsi, il faut attendre le XVI^{e} siècle siècle pour voir apparaître un dictionnaire de référence (le dictionnaire français-latin de Robert Estienne[21]) [Bea27], et la normalisation de la graphie mettra encore quelques siècles avant de se généraliser.

Le formalisme orthographique n'est donc pas consubstantiel à l'écrit, il s'agit d'un ajout, souvent tardif, volontaire, qui paraît intervenir principalement lorsque des écrits sont largement diffusés (succédant à l'invention de l'imprimerie pour le français, au développement de la bureaucratie pour le latin ou le chinois, et jamais attesté pour les graphies « magiques » réservées à un petit cercle d'initiés).

Beaucoup de phénomènes graphiques relevés en tchat peuvent ainsi s'interpréter comme relevant d'un écrit informel, nullement étrange ou exotique. Par exemple, l'insertion de signes autonomes, porteurs d'une signification propre, dans des mots, d'après leur valeur phonétique et sans tenir compte de leur valeur sémantique ; ainsi les formes « 2m1 », « 2main » et « dem1 » (voir nombre d'occurrences en table 1.15), pour « demain », font-elles directement écho à notre exemple du maya.

Graphie	Nombre d'occurrences	Pourcentage
2m1	71	46%
2main	7	5%
dem1	77	50%
Total	155	100%
demain	4967	

Table 1.15 – Nombre d'occurences de diverses graphies de « demain » (employées dans ce sens) sur un échantillon représentatif de de 1 860 085 messages (9 950 821 mots) extrait de notre corpus.

On peut aussi relever des graphies résultant d'une créativité sémantico-graphique, telles que « Micro$oft » pour « Microsoft »[22], dans lesquelles on insère un signe possédant des valeurs lexicale et sémantique propres, qui sont cette fois toutes deux conservées dans la graphie ainsi formée.

[21] *Dictionarium latino-gallicum, contenant les motz et les manières de parler françois, tournez en latin.*

[22] 14 occurrences dans dans notre corpus

De ce point de vue, l'usage d'une graphie « phonétique » n'a elle non plus rien de surprenant.

Enfin l'usage plus ou moins codifié d'abréviations n'a rien d'exceptionnel. Comme le fait remarquer [GdNV04], les abréviations sont utilisées depuis l'antiquité, et certaines on d'ailleurs été consacrées par la norme, à l'image de la cédille (au départ un caractère « z »), ou de l'*Umlaut* allemand (à l'origine un caractère « e »).

S'agissant d'une forme d'écrit éphémère, spontanée et surtout à diffusion restreinte, il n'est donc pas étonnant que le tchat partage des caractères avec des graphies peu formelles.

2.2.6 Étude quantitative des divergences graphiques

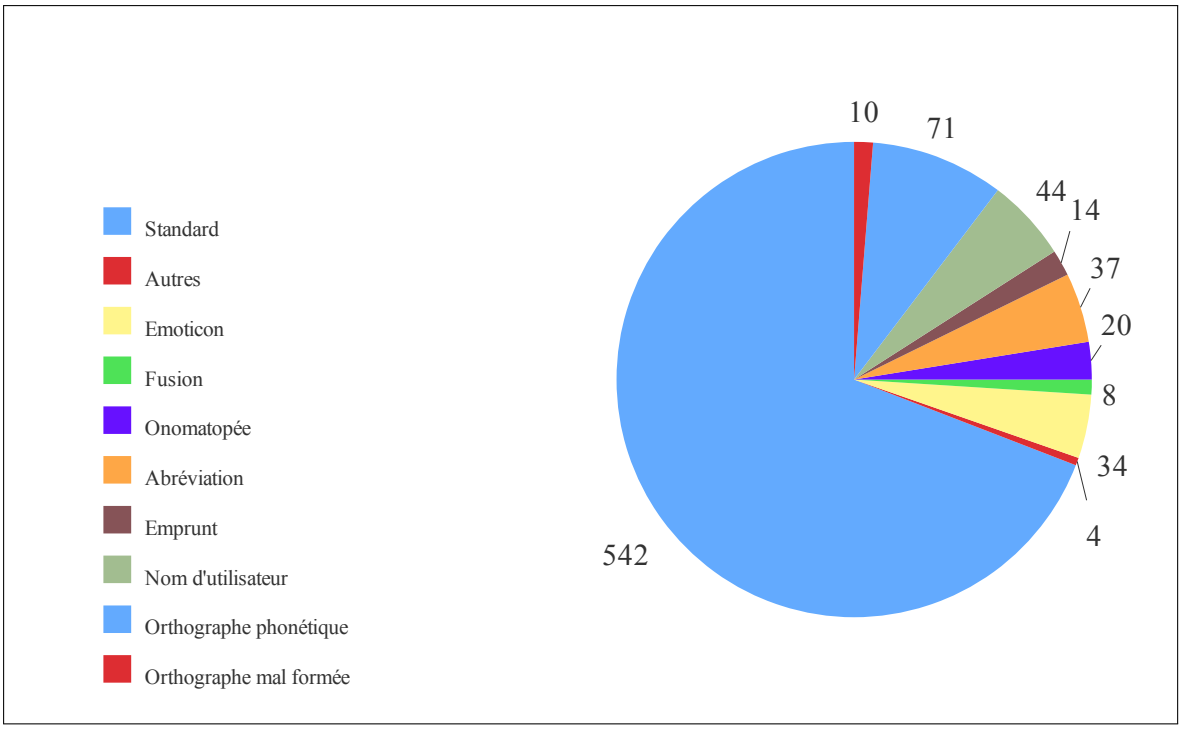


Figure 1.19 – Phénomènes lexicaux relevés dans 200 énoncés (783 mots) sélectionnés aléatoirement sur le canal #18-25ans.

La figure 1.19 donne une idée générale de l'importance de ces divergences graphiques (on ne considère pas les problèmes d'accord, mais si les graphies correspondent à une graphie existante dans l'écrit standard). On constate que les deux tiers environ des « mots » (caractères entre deux espaces, y compris émoticons) sont correctement orthographiés (au sens défini précédemment, c'est à dire sans tenir compte des accords), et que les mots involontairement mal orthographiés (« orthographe mal formée ») sont rares. Par contre, un nombre significatif de mots relèvent de la graphie phonétique (« orthographe phonétique ») décrite précédemment. Le

recours aux abréviations, aux émoticons, aux onomatopées, ainsi que les références aux autres utilisateurs, se révèlent aussi assez courants. Enfin, on relève quelques cas d'emprunts (des anglicismes en l'occurrence) et de fusions de mots (« jme demande », « jte dis », « ça mva », etc.).

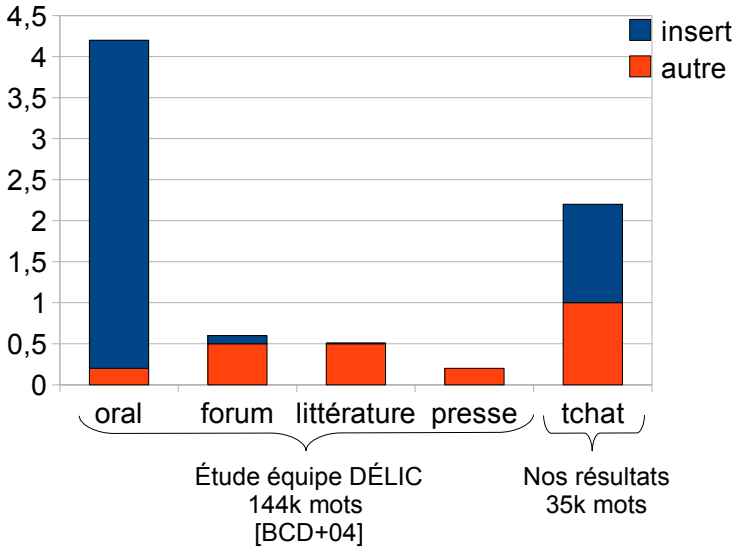


Figure 1.20 – Usages de « bon » comme insert dans 4 corpus de 440 000 mots [BCD+04], et un extrait de 35 000 mots issu du canal #18-25ans.

D'un point de vue plus général, il semble qu'un corpus de tchat comporte nettement plus de formes qu'un corpus écrit « standard » ou oral équivalent, quand on compare notre corpus (figures 1.14 et 1.15) avec ceux décrits dans [GAD02]. Toujours par comparaison avec les corpus écrit et oral de [GAD02], on peut constater que le nombre de formes présentes plusieurs fois, par rapport au nombre total de formes, est beaucoup plus faible en tchat (35% des formes pour le canal #18-25ans) que dans l'écrit « standard » (62%) et l'oral (70%). Cela peut s'expliquer par la fréquence de ces divergences graphiques.

La langue du tchat est souvent comparée à la langue orale, mais sans mesure quantitative. Nous avons comparé ce qui est selon l'équipe DELIC [BCD+04] un bon indice de l'oralité : l'emploi de « bon » en insert, c'est à dire employé comme interjection plutôt que comme adjectif.

La figure 1.20 présente les résultats obtenus. [BCD+04] voit dans le léger excédent d'emploi de « bon » en tant qu'insert dans les forum le signe que « *les nouvelles formes de communication écrite ont tendance à reprendre certaines caractéristiques du langage parlé* », mais force est de constater que ces marques d'oralité sont nettement plus prononcées en tchat. On peut constater que selon cet indice, le tchat se trouve à mi-chemin entre écrit et oral à deux titres :

(1)	**oui mais** c'est pas définitif **hein** madame votre question
(2)	le problème c'est que **bon** ça ça stoppe + quand on écarte les bras
(3)	en pleine campagne trouver des choses à faire **ben** c'est devenir agriculteur ou éleveur **quoi**
(4)	c'était un système d'habitude qui s'était instauré une routine **tu vois** + et j'étais pas amoureux
(5)	on n'avait pas pas amené au magasin le petit bout de tapisserie **tu sais** pour voir la couleur

Table 1.16 – Exemples d'inserts (en gras) dans un corpus de parole [Fal05]).

1. la fréquence d'emploi intermédiaire de la forme « bon » elle-même,
2. la proportion intermédiaire d'emploi de « bon » en tant qu'insert.

On notera qu'un autre outil de communication médiée par ordinateur, le forum, produit des énoncés nettement plus proches de l'écrit traditionnel ; d'où l'importance de se garder de tout amalgame entre ces différents types de communication.

Conclusion

L'écrit tchaté est parfois présenté comme quelque chose d'exotique, un produit étrange de l'ère numérique. Quelques études qualitatives antérieures à la thèse, notamment [Pie03b] et [Pie03a], complétées par nos études quantitatives, ont montré qu'il n'en était rien. Les divergences observées par rapport à l'écrit standard s'expliquent par le caractère spontané, non planifié, du tchat, par des jeux de langue et un effet de communauté.

En ce qui concerne le caractère non planifié du tchat, il convient de relativiser le parallèle que l'on pourrait poser entre tchat et parole spontanée : les productions de tchat présentent toujours un peu plus de planification que la parole spontanée, dans la mesure où les messages de tchat sont composés, et éventuellement corrigés, *avant* d'être volontairement envoyés dans le dialogue par l'utilisateur. Cette phase obligatoire de planification peut être très réduite, dans le cadre de discussions informelles mettant l'accent sur l'interactivité, mais elle existe néanmoins.

Par ailleurs, comme le notait [GdNV04], on retrouve des phénomènes graphiques communs aux graphies peu normées. Au niveau de la syntaxe, le tchat emprunte beaucoup à l'oral informel. Mais étant utilisé dans le cadre d'un dialogue en « temps réel », il diverge des autres graphies peu normées, et emprunte aussi des traits aux formes de productions langagières dans lesquelles planification et production sont quasiment synchrones, et en premier lieu l'oral spontané. Enfin, étant utilisé dans le cadre d'un dialogue, il introduit des éléments verbaux et surtout paraverbaux indispensables à la vérification de l'état du canal et au souci de congruence/empathie.

Les utilisateurs de tchat sont conscients de problèmes liés à l'écrit tchaté, et en particulier concernant le caractère non standard de cette forme d'écrit. Les tchateurs s'entraident, et se corrigent lorsqu'ils l'estiment nécessaire.

En terme de fonctionnalités pour un outil de tchat en langue seconde, on voit l'importance de l'écrit non-standard pour les tchateurs : loin d'être un phénomène purement involontaire et subi, il est souvent volontaire et motivé par des considérations d'ordre social, afin d'enrichir l'écrit standard d'une dimension phatique/paraverbale. Un outil d'aide au tchat, en langue seconde ou non, devrait donc être capable de fonctionner avec de l'écrit non standard.

De plus, on remarque que les tchateurs corrigent parfois leurs messages *a posteriori*. Actuellement, les outils de tchat ne proposent pas cette fonctionnalité, et les utilisateurs doivent se débrouiller pour exprimer cette correction dans un message ultérieur. Il serait utile de fournir une telle fonctionnalité aux utilisateur de notre système de tchat en langue seconde.

Nous verrons, dans la partie suivante, ce qui se passe lorsque l'on passe à un dialogue en *langue seconde*.

Chapitre 2

Le dialogue en langue seconde : spécificités, outils

La Dispute du Saint-Sacrement, Raphaël, XVI^e^ siècle. En Europe, durant le Moyen-Âge et la Renaissance, les débats théologiques et scientifiques, ou *disputes*, se tiennent en latin, entre locuteurs dont ce n'est pas la langue maternelle.

1 Le multilinguisme

1.1 Terminologie

Certains termes sont utilisés dans un sens particulier dans le cadre de l'étude du multilinguisme :

Langue seconde désigne toute langue qui n'est pas maternelle.

Locuteur natif (*native speaker* ou *NS*) désigne un locuteur s'exprimant dans sa langue maternelle.

Locuteur non natif (*non native speaker* ou *NNS*) désigne un locuteur s'exprimant dans une langue seconde.

Dialogue en langue seconde désigne un dialogue dans lequel au moins l'un des locuteurs est non natif.

1.2 Histoire et situation actuelle

Les situations de multilinguisme, avec la coexistence de langues véhiculaires[1] et de langues vernaculaires[2], sont sans doute aussi vieilles que le langage. Pour ce qui est des temps historiques, les exemples sont nombreux, et nous nous contenterons de quelques cas européens :

grec *koinè* [3] langue de prestige des lettrés du bassin méditérannéen, puis des soldats d'Alexandre, parlée par les diplomates et marchands de l'Antiquité (IV^{e} siècle avant-IV^{e} siècle après J.C.). Cette langue restera celle de la médecine durant tout le Moyen-Âge.

latin médiéval, puis humaniste langue de l'enseignement et de la liturgie, dans toute l'Europe (IX^{e} siècle-XX^{e} siècle).

lingua franca langue des marchands, marins, bagnards et esclaves de l'ensemble du bassin méditerranéen (XI^{e} siècle-XIX^{e} siècle siècle) ; c'est un pidgin[4] de langues romanes (castillan, catalan, occitan, français, dialectes italiens), avec des emprunts à l'arabe, à l'hébreu, au maltais, au turc, etc. plus ou moins prononcés suivant les régions et les locuteurs.

Toutes ces langues véhiculaires se caractérisent par une orthographe peu normalisée, et une simplification de la phonologie, de la morphologie, de la syntaxe et du lexique.

Aujourd'hui encore, il est courant de maîtriser au moins sommairement une langue en plus de sa langue maternelle [Lec]. Ainsi, l'hindi est la langue maternelle de 18% des indiens, mais est compris à divers degrés par 50% de la population indienne. En Afrique orientale, le swahili est presque exclusivement une langue véhiculaire, parlée par 40 à 50 millions de personnes.

Au niveau planétaire, l'anglais, ou plus exactement sa forme simplifiée, le *globish*, est souvent présenté comme la *lingua franca* globale du XX^{e} siècle. Des dialectes spécialisés existent, comme l'*airspeak* (dialecte des pilotes et contrôleurs aériens), le *seaspeak* (dialecte des marins), etc.

Une étude récente de l'Union Européenne (2005 - cf. figure 2.1) montre que 56% des Européens parlent au moins une langue étrangère. Un peu plus de la moitié de la population de l'Union maîtrise l'anglais, et pour une majorité en tant que langue étrangère, une caractéristique qui se retrouve aussi au niveau mondial : il y a plus de locuteurs de l'anglais non natifs que de natifs.

[1] Une *langue véhiculaire* est une langue, souvent simplifiée, servant de moyen de communication entre populations de langues différentes. Elle s'oppose à la langue vernaculaire, parlée localement par une population (Wikpedia).

[2] On appelle *langue vernaculaire* la langue locale communément parlée au sein d'une communauté. Ce terme s'emploie souvent en opposition avec le terme langue véhiculaire (Wikpedia).

[3] *κοινή* : « commun ».

[4] Le terme de *pidgin* désigne différentes langues véhiculaires simplifiées créées sur le vocabulaire et certaines structures d'une langue de base (Wikipedia). Un pidgin plus structuré, devenu vernaculaire, sera parfois qualifié de *créole*.

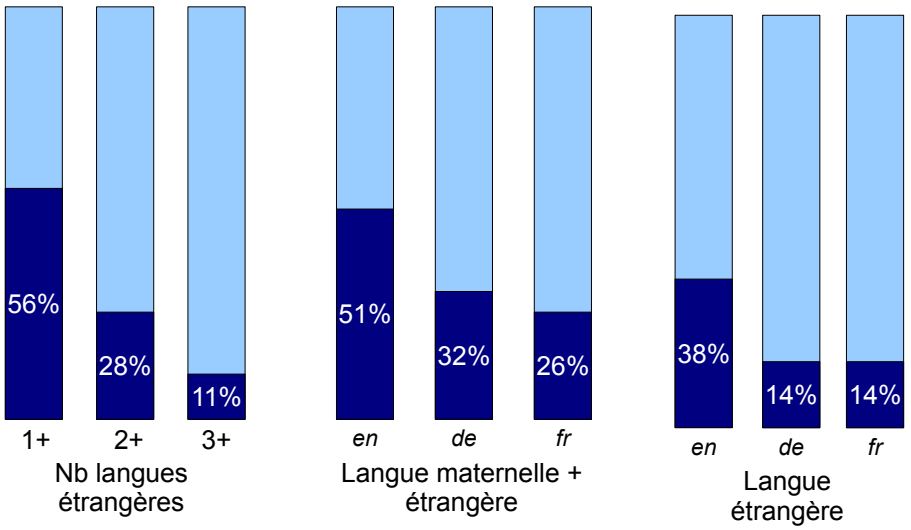


Figure 2.1 – Les langues étrangères dans l'UE à la fin-2005. Étude portant sur 28 694 personnes de l'UE27 + Turquie et Croatie [lgu06].

1.3 Des situations variées

Aujourd'hui, le multilinguisme se développe selon plusieurs axes. Tout d'abord, on voit émerger de nombreux ensembles supranationaux (et supralinguistiques), par exemple l'Organisation des Nations Unies, où seules 6 langues ont un statut officiel (anglais, arabe, espagnol, français, mandarin, russe), ou encore l'Union Européenne, où au contraire un maximum de langues ont un statut officiel (23 langues en 2009). L'Inde est un autre exemple important d'ensemble supralinguistique (au niveau fédéral, seuls l'hindi et l'anglais sont reconnus, alors qu'au niveau local le pays compte officiellement 22 autres langues, et encore 13 autres langues non officielles comptant plus de 5 millions de locuteurs).

Dans le cadre de sociétés multinationales ou de projets internationaux (recherche scientifique, projets informatiques open-source), des individus de langues maternelles différentes sont amenés à travailler en collaboration. Le tourisme de masse est une autre donnée importante. En 1950, on comptait 25 millions de touristes internationaux, en 2000 ils étaient 500 millions. Hors tourisme, on estime actuellement à 4% de la population mondiale (soit environ 240 millions de personnes) le nombre d'expatriés. Enfin, Internet réduit les barrières spatiales entre les communautés linguistiques, et amène des individus à dialoguer dans une langue qui n'est pas leur langue maternelle. Il faut donc compter avec un grand nombre de situations et de besoins.

Schématiquement, le locuteur non natif peut se définir selon plusieurs critères :

Maîtrise du domaine, axe qui n'est pas spécifique au dialogue en langue étrangère.

Expertise linguistique, de la langue véhiculaire ; on distingue les locuteurs de langue seconde des locuteurs de langue étrangère [Kac92].

Langue maternelle, qui influence d'autant plus ses productions que son expertise linguistique est faible [Gus99].

Expertise culturelle, qui peut l'amener à interpréter plus ou moins correctement les attitudes et les intentions de son interlocuteur [Del06] ; par exemple les scripts de politesse, ou plus spécifiquement l'usage de l'impératif en allemand pour un anglophone, ou encore la différence de valeur du modal *must*/*müssen* dans ces deux langues.

En pratique, on distingue surtout les dialogues symétriques (les deux locuteurs sont non natifs) des dialogues asymétriques (l'un des locuteurs est natif).

Les finalités peuvent être très diverses, comme pour le dialogue en langue native, auquel vient s'ajouter le dialogue d'apprenant, où le dialogue est souvent à lui-même sa propre finalité.

2 Le dialogue en langue étrangère

Le dialogue non natif a été étudié sous différents angles :

Enseignement des langues : c'est de loin l'approche la plus répandue. Elle s'intéresse en premier lieu à la grammaticalité des énoncés, aux « erreurs » typiques par rapport à la norme de la langue native, et aux moyens d'y remédier.

(Anglais) langue internationale [5] : étudie l'anglais employé en tant que langue étrangère. Cette approche met l'accent sur l'efficacité de la communication. Dans cette perspective, on ne pose plus la langue native comme standard à atteindre par les locuteurs ; on s'intéresse à l'outil de communication interlingue. D'autres langues peuvent être étudiées suivant cette approche, mais en pratique c'est presque toujours l'anglais qui fait l'objet d'études.

Interculturalité : c'est une approche plutôt sociologique visant à déterminer l'influence d'une langue/culture donnée sur une autre.

Par la suite, on s'intéressera principalement à l'aspect «langue internationale». On ne peut bien entendu totalement disjoindre ces points de vue : en s'exprimant, tout locuteur non natif va voir ses compétences progresser, et sa culture peut être déterminante pour certains scripts.

2.1 Description

Dans le domaine de l'anglais langue internationale, on peut distinguer les travaux récents de Christiane Meierkord [Mei96] [Mei00] et Barbara Seidlhofer [Sei03] [HBS08]. Le dialogue non natif se caractérise principalement par un lexique pauvre, une co-construction des énoncés nettement plus développée qu'en langue native, et l'importance de l'interaction, des marques d'attention, des retours.

[5] *(English) as lingua franca.*

D'un point de vue morphologique, les langues sont souvent simplifiées lorsqu'elles sont employées dans une fonction véhiculaire [Sei05]. Par exemple l'anglais non natif se caractérise par l'absence totale de flexion personnelle du verbe (*I like*, *She like*), l'usage interchangeable des pronoms *who* (animé) et *wich* (inanimé), l'omission ou au contraire la surgénération d'articles, la flexion de mots invariables (*informations*, *advices*), etc. De plus, afin d'accroître l'intelligibilité des énoncés, les locuteurs ont tendance à ajouter des prépositions (*discuss about something*, *phone to somebody*) et des substantifs (*how long time*).

La langue utilisée dans le cadre du dialogue non natif n'est pas nécessairement fixée au préalable. Elle peut faire l'objet de négociations entre les locuteurs, et même changer en cours de dialogue [Mon05].

2.1.1 Difficultés pour le locuteur natif

En ce qui concerne les difficultés, elles ne sont pas uniquement l'apanage du locuteur non natif et de la méconnaissance de la langue standard. En effet, il ne s'agit pas d'un dialogue en langue standard : le locuteur natif peut avoir du mal à faire abstraction de sa connaissance des idiomatismes. Cela peut poser problème [Gnu00], en production (il doit faire l'effort de produire des énoncés pauvres en idiomatismes, et déterminer lesquels sont susceptibles d'être connus de son interlocuteur), mais aussi en réception, lorsqu'une locution, produite de manière naïve par le locuteur non natif, est interprétée différemment, de manière intuitive et immédiate, par le locuteur natif (cf. table 2.1, ci-après). Ce problème est appelé « idiomatisme unilatéral[6] » [Sei04] par les auteurs.

2.1.2 Difficultés pour le locuteur non natif

En production, les locuteurs non natifs souffrent de la difficulté d'approfondir les énoncés. Dans l'exemple 2.1, un interviewer essaye de « faire parler » un étudiant en première année de français, ayant déjà étudié le français dans le secondaire. Comme on peut le constater, c'est assez laborieux : l'étudiant est incapable de développer un sujet, et ne fait que répondre le plus succinctement possible aux questions. De fait, on caractérise aussi le dialogue non natif par la pauvreté de son lexique, les hésitations des locuteurs et leurs incertitudes quant à la correction de leurs énoncés (surtout vis à vis de locuteurs natifs) et par conséquent, sa superficialité.

On constate aussi, chez cet étudiant, une tendance à s'exprimer par enchaînement d'énoncés simples (dans l'exemple 2.1, séquences (2) - (4) , (6) - (7)). Plutôt que de construire des structures syntaxiques complexes, il privilégie une succession de structures simples ; un type d'enchaînement que l'on retrouve à l'écrit en tchat, même monolingue. De fait, une syntaxe simplifiée peut se révéler relativement productive, mais en production, le locuteur non natif se retrouve souvent coincé par la pauvreté de son lexique, et de l'utilisation de ce lexique (usage, collocations).

[6] *unilateral idiomaticity.*

(1)	Loc1	what are you mumbling ?
(2)	Loc2	that song - hhhhm. hh. what's that woman's name ? that singer - *the chocolate woman* !
(5)	Loc1	areta franklin
(6)	Loc2	no : ! she's not black !
(8)	Loc1	you said she was black !
(9)	Loc2	yes you did !
(10)	Loc1	yes you did. you said that chocolate woman.
(11)	Loc2	yes, the one that sings that your love is better than chocolate
(12)	Loc1	sarah maclofn ? she's white
(13)	Loc2	I never said she was black !
(14)	Loc1	yes - you - did
(15)	Loc2	NO I didn't ((frustrated))
(16)	Loc1	((laughs))
(17)	Loc2	whats so funny ? ((quite angry))
(18)	Loc1	honey, a chocolate woman IS a black woman.
(19)	Loc2	oh - how am I supposed to know that ?

Table 2.1 – Dialogue dans un couple bilingue. Loc1 est locuteur natif, Loc2 non natif [GL05].

2.2 Incompréhensions et coopération

En réception, se posent des problèmes d'incompréhension. Comme nous l'avons vu en première partie, l'incompréhension n'est pas un phénomène spécifique au dialogue non natif, et les sous-dialogues de clarification non plus. Ce phénomène a aussi une part culturelle ; par exemple, pour l'étude du japonais, on utilise souvent une unité au-dessus du tour de parole, le *wadan*, pour désigner une opération de coconstruction du sens, impliquant plusieurs tours [Pol93].

Il est néanmoins très fréquent en dialogue non natif : selon Oviatt et Cohen [OC91], ces sous-dialogues de clarification occupent jusqu'à 30% des tours de parole avec interprète humain. Varonis et Gass [VG85] ont montré qu'un dialogue natif/non natif comporte plus de séquences de clarification qu'un dialogue natif/natif, et qu'un dialogue non natif/non natif en comporte plus qu'un dialogue natif/non natif.

Marie-Luise Pitzl [Pit05] a mené une étude sur les incompréhensions en anglais langue internationale. L'incompréhension peut être partielle ; les locuteurs tentent alors de maintenir un certain degré de compréhension tout au long du dialogue. Elle reprend un modèle qui situe les signes d'incompréhension de l'interlocuteur le long d'un continuum [VBR96], des signes les plus implicites (« symptômes ») aux plus explicites (« signaux ») :

- ignorer le passage problématique
- pause avant de répondre
- réduction des signes phatiques

(1)	Loc1	pourquoi est ce que tu as choisi l' université de Salford ?
(2)	Loc2	um j' aime Manchester.
(3)	Loc2	et je *trust* euh j' ai voulu étudier à Manchester et à Salford.
(4)	Loc2	j'aime le cours.
(5)	Loc1	tu viens de quelle partie de l'Angleterre ?
(6)	Loc2	um pas d' Angleterre.
(7)	Loc2	du pays de Galles.
(8)	Loc1	du pays de Galles ?
(9)	Loc2	oui.
(10)	Loc1	euh d' ou au pays de Galles ?
(11)	Loc2	euh le sud.
(12)	Loc1	le sud ?
(13)	Loc2	oui de Swansea.
(14)	Loc1	um est ce que tu as le mal du pays ici ?
(15)	Loc2	um oui beaucoup.
(16)	Loc1	tu rentres pour Noël ?
(17)	Loc2	euh oui.
(18)	Loc1	ok où est ce que tu habites à Salford ?
(19)	Loc2	euh j' habite à halls.
(20)	Loc1	halls ?
(21)	Loc2	halls.

Exemple 2.1: Étudiant anglais, ayant étudié le français dans le secondaire et en première année de *Modern Languages* à l'université de Salford. Corpus Salford [MM].

- formulation d'hypothèse
- reprise du passage problématique
- questions
- commentaires métalinguistiques

L'efficacité et le coût de ces différents procédés croissent en fonction de leur caractère explicite. Par conséquent, les locuteurs alternent les différents procédés, dans une tentative de maintenir à la fois un niveau de compréhension acceptable et un dialogue fluide.

Les causes d'incompréhension sont multiples, mais il s'agit principalement de causes linguistiques et pragmatiques (liées à des divergences culturelles) [GL05]. En ce qui concerne les causes linguistiques, elles sont principalement lexicales.

Les réparations peuvent être commencées par les locuteurs natifs, aussi bien que non natifs [GL05]. Ils font généralement preuve d'une grande coopération [Mei96] [Fir96], si bien que ces incompréhensions sont presque toujours résolues. Lorsqu'un locuteur natif est présent, c'est souvent lui qui clôt les sous-dialogues de clarification [GL05].

La négociation du sens qui s'ensuit a été formalisée par [VG85] (cf. figure 2.2), dans un modèle comprenant deux parties principales : le *déclencheur*, qui est l'indice

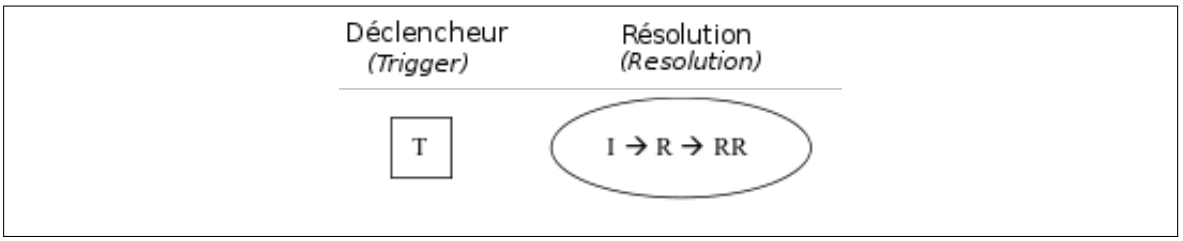


Figure 2.2 – Modèle de négociation du sens [VG85]

amenant le locuteur à penser qu'il a été mal compris, et *résolution*, qui se décompose en identification du problème (I), réponse (R) et réaction à la réponse (RR). Cette négociation peut être extrêmement simple. Ainsi, dans l'exemple qui suit (table 2.2), un problème est identifié (I) au tour (9) , corrigé (R) en deux mots au tour (10) , et cette réponse est confirmée de façon succincte au tour (11) (RR). À cette

(1)	Loc1	and we er : developed (.) (to) (.) store fit er size rack. er which er (.) i er : showed the
(2)	Loc4	mhm
(3)	Loc1	pictures on page twenty-eight (.) and also we called it wire rack (.) so that was a bi- er : that was our er <spel> a </spel> and <spel> p </spel> er
(4)	Loc2	<soft> seventeen </soft>
(5)	Loc1	page seventeen ?
(6)	Loc2	<soft> mhm </soft>
(7)	Loc1	oh yeah pa- sorry page seventeen
(8)	Loc5	mhm (.)
(9)	Loc1	we produced about three hundred of that
(10)	Loc4	of those ?
(11)	Loc2	<soft> yeah </soft>

Table 2.2 – Dialogue entre cinq locuteurs, dont deux apprenants (2 : coréophone, et 5 : germanophone), à propos d'images présentes dans un dossier qui leur a été remis [Pit05].

séquence I-R-RR viennent souvent s'ajouter des remerciements, des excuses, etc., avant le retour au dialogue [GL05].

2.3 Dialogue écrit en langue seconde

On peut noter que le tchat et le dialogue en langue seconde convergent en ce qui concerne la longueur des tours de parole, qui ont tendance à être courts. On peut aussi retrouver, pour des raisons différentes, le goût pour les structures syntaxiques simples. Par contre, un tchat en langue seconde sera nécessairement plus normé qu'un tchat en langue première. Toutefois, nous avons vu que les graphies peu normées pouvaient évoluer vers une forme plus normée lorsque les circonstances le nécessitent, et cela ne devrait donc pas poser de problème en tchat, pourvu que l'on en donne les moyens aux utilisateurs.

Peu de données sont disponibles quant à l'usage du tchat en langue seconde, et les études sur le sujet portent principalement sur l'utilisation pédagogique du tchat dans l'enseignement des langues (voir par exemple les travaux de Susana Sotillo [Sot05] [Sot06]). Néanmoins, les quelques dialogues accessibles dans ces études ainsi que les quelques cas relevés dans notre corpus du français confirment cette hypothèse.

Conclusion

Le multilinguisme est un phénomène ancien, et largement répandu. De nombreuses régions du monde ont vu et voient toujours cohabiter plusieurs langues sur un même territoire. Un dialogue qui implique un locuteur non natif n'est pas toujours monolingue, les locuteurs peuvent changer de langue au fil de la conversation, voire parler deux langues différentes. La langue utilisée est simplifiée, à des degrés divers dépendant du niveau du participant le plus faible, au niveau duquel les locuteurs plus compétents vont essayer de s'abaisser. La production linguistique qui en découle est généralement lexicalement pauvre, avec une syntaxe simplifiée, alternant des tours de parole courts. Les locuteurs sont très demandeurs de retours, s'entraident et s'encouragent mutuellement bien plus que dans une conversation entre locuteurs natifs, se rapprochant des comportements observés en langue native dans des situations bruitées ou de réparation ; le paraverbal et le non-verbal deviennent essentiels. Les locuteurs disposent de divers procédés pour signaler les problèmes éventuels, et réparer le dialogue, les moins coûteux étant aussi les moins puissants. Ils vont chercher à maintenir un bon rapport compréhension/fluidité dans le dialogue.

3 Étude de l'existant

3.1 Les livres de phrases électroniques

Les livres de phrases électroniques sont des appareils comportant des listes de phrases écrites, alignées en plusieurs langues. L'utilisateur peut choisir une phrase, et obtenir sa traduction dans la langue de son choix.

Il existe de nombreux systèmes de ce type, plus ou moins évolués. Nous n'en citerons que quelques-uns, les plus représentatifs de l'état de l'art en ce domaine.

3.1.1 Phraselator

L'un des plus simples est le *Phraselator*[7] de la société *VoxTec*. Il s'agit d'un petit appareil de la taille d'un PDA, initialement développé pour l'armée américaine, qui comporte des listes de phrases en anglais, alignés sur des fichiers audio correspondants.

L'utilisateur prononce la phrase dont il souhaite obtenir la traduction, le système la recherche dans la liste et émet la traduction correspondante. Une recherche textuelle de phrase est aussi possible (cf. figure 2.3).

[7] `http://www.voxtec.com/phraselator/`

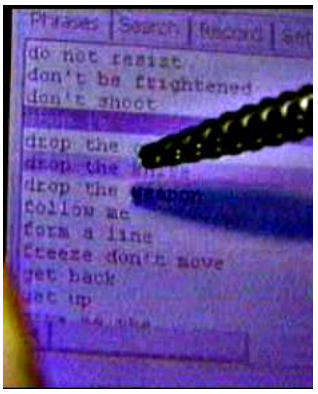


Figure 2.3 – L'interface du livre de phrase digital *Phraselator*

Le système n'est pas prévu pour permettre à l'auditeur non-anglophone de répondre. Il peut néanmoins enregistrer les réponses, qui pourront être soumises ultérieurement à un interprète pour traduction.

3.1.2 Interpreter

Sur PDA, *Interpreter*[8], de *SpeechGear*, est l'un des plus complets. Le logiciel est pilotable à la voix (en option), et une synthèse vocale est disponible en sortie. Il est bidirectionnel (l'interlocuteur peut choisir une réponse dans sa langue), et permet les variables lexicales . Les phrases sont classées par situation (restaurant, etc.).

Certaines de ces variables peuvent être contraintes par une classe , ou bien libres (cf. figure 2.4). Les variables sont instanciées dans la phrase, qui peut subir des modifications morphologiques (par exemple accord en genre et en nombre de l'article en allemand).

De plus, chaque phrase peut correspondre à plusieurs traductions, que l'utilisateur peut choisir en fonction de la situation. On peut choisir le degré de politesse (distance, familiarité, respect, etc.), si l'on s'adresse à un groupe ou bien à individu seul, à un homme ou une femme.

3.1.3 PSP Talkman

Enfin *Talkman* de *Sony*, un logiciel pour console de jeu portable *PSP*, se veut beaucoup plus ludique (cf. figure 2.5). Les phrases sont organisées en scénario, et en dehors d'un système de « favoris », il n'est pas possible de panacher. Le logiciel peut se piloter à la voix, les phrases sont alors réorganisées en fonction de leur ressemblance avec le résultat de la reconnaissance vocale.

[8] `http://speechalator.com/interprete.aspx`

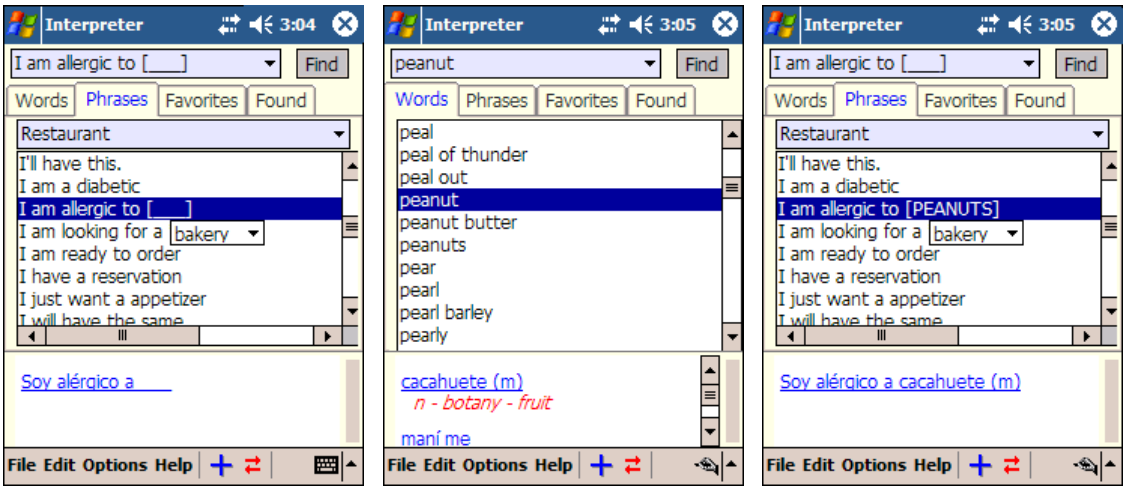


Figure 2.4 – Interface du livre de phrase digital *Interpreter*

Figure 2.5 – Interface du livre de phrase digital *Talkman*

3.1.4 Étude sur les livres de phrases digitaux : SurviTra

Dans le cadre du projet franco-indien CIFLI-SurviTra (Communication sur Internet en Français et Langues Indiennes - Survival Translation kit, programme ARCUS-Inde) [BBB+07] [FFG09], nous avons développé une plate-forme en vue d'une approche de TA par pivot, à l'aide de composants UNL.

UNL (Universal Networking Language) est un formalisme lexico-sémantique destiné à exprimer tout énoncé textuel d'une langue naturelle de manière non ambiguë, à l'aide d'un graphe (cf. figures 2.6 et 2.7). Les unités lexicales, portées par les nœuds des graphes, sont appelées des UW (*Universal Word*). Des composants de déconversion (graphe UNL vers langue) existent pour une douzaine de langues, mais l'enconversion (langue vers graphe UNL), plus complexe, est souvent disponible plus tard, et avec une couverture partielle, évolutive. Ainsi, en ce qui concerne les langues

impliquées dans le projet, la déconversion du français est actuellement disponible, à environ 70% de couverture sur le domaine ; l'opération inverse (enconversion du français) est réalisée à 20% de couverture syntaxique, sur des modèles de phrases simples d'abord testés sur le domaine. La déconversion du hindi est également en cours de développement [Bla05] [SDV+07], mais l'enconversion du hindi est en projet.

Pouvez-vous recommander un bon restaurant pour le déjeuner ?
क्या आप मुझे दोपहर भोजन के लिए एक अच्छा होटल बता सकते है?
Can you recommend a good restaurant for lunch?

Figure 2.6 – Exemple de phrase en version trilingue (polyphrase).

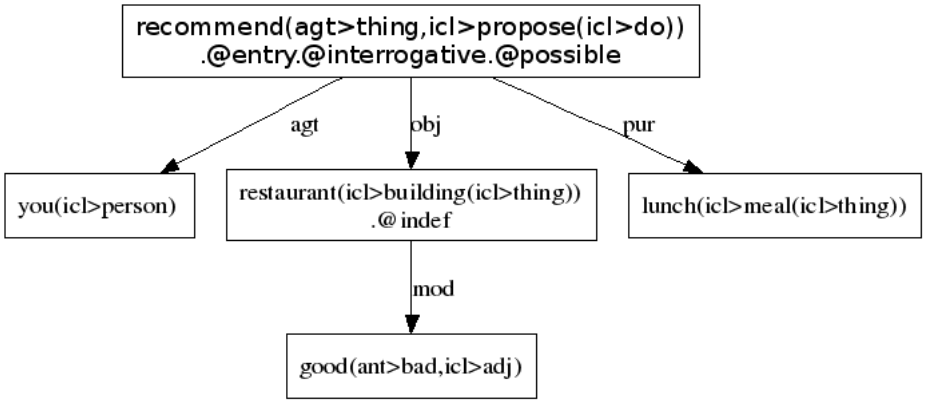


Figure 2.7 – Graphe UNL correspondant.

Les composants UNL étant à des stades très divers de développement, une architecture modulaire s'est imposée. Dans un premier temps, le système s'appuie sur des phrases alignées, comme les livres de phrases classiques. Cela permet d'étudier les aspects ergonomiques du système, indépendamment de l'avancement des composants UNL.

Ensuite, au fur et à mesure de leur développement, des composants UNL (enconvertisseurs/déconvertisseurs) sont ajoutés. En fonction de l'avancement des enconvertisseurs, il est probable que seule une fraction des phrases soit disponible au format UNL. Dans le cas où il ne peut pas obtenir une phrase par déconversion, le système utilise les phrases alignées à la place. Le système permet des variables lexicales, avec ou sans UNL.

Aucun traitement morphologique n'est réalisé au niveau du système : outre l'approche pivot, l'intérêt d'UNL est d'intégrer la question de l'adaptation morphologique au niveau de la déconversion. Il est en effet aisé de concevoir un « UNL à variables lexicales », c'est à dire un graphe en pseudo-UNL, dans lequel on a seulement à instancier une variable par une unité lexicale pour obtenir un graphe UNL valide, sans se soucier de sa morphologie en langue.

Ainsi, pour l'exemple de la figure 2.7, on peut remplacer l'UW « *lunch* » - `lunch(icl>meal(icl>thing))` par une variable lexicale `[$MEAL]`, qui incorpore les UW `breakfast(icl>meal(icl>thing))`, `lunch(icl>meal(icl>thing))` et `dinner(icl>meal(icl>thing))`.

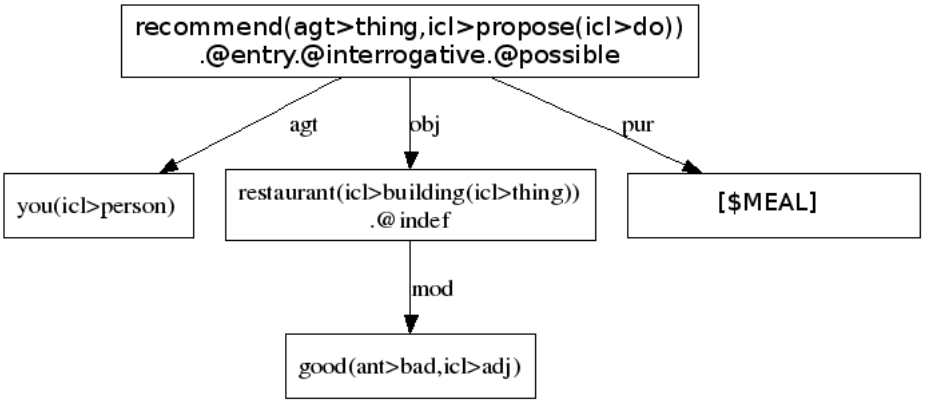


Figure 2.8 – Graphe UNL avec une variable lexicale.

Suivant l'approche classique des livres de phrases, les phrases sont organisées en domaines à plusieurs niveaux (sous-domaines, etc.). Les variables lexicales correspondent à des classes de mots/UW fermées. Pour permettre de distinguer certaines ambiguïtés, dans le cas où un même terme dans une langue peut se traduire de plusieurs manières dans une autre (par exemple « bonjour » qui peut se traduire en anglais par « *good morning* » ou « *good afternoon* » selon le moment de la journée), chaque entrée lexicale se voit associer un champ « Précision ». Cela est aussi valable pour les phrases, qui peuvent changer selon le degré de politesse, etc. Enfin, pour les termes culturellement spécifiques, fréquents dans le domaine de la restauration (par exemple, « lassi » se traduira par « lassi » en français, ce qui risque de ne pas beaucoup avancer l'utilisateur ; il pourra apprécier de savoir qu'il s'agit d'une boisson à base de yaourt ou de lait fermenté), les lexies se voient attribuer un champ supplémentaire d'explication/encyclopédie.

fra (entrée)	fra (précision)	fra (explication)	hin (entrée)
gulab jamun		boules de farine dans un sirop sucré	गुलाब जामुन
lassi		boisson à base de yahourt fermenté	लस्सी
bonjour	après-midi		गुड ऑफ्टरनून
bonjour	matin		सुप्रभात

Table 2.3 – Exemples de précisions et d'explications pour quelques entrées en français.

En ce qui concerne l'interface, plusieurs approches sont à l'étude. Un premier prototypage SurviTra-0.2 a permis d'explorer l'hypothèse de communication Web entre deux locuteurs conversant en configuration mono-machine ou depuis deux postes. Il a préparé le traitement de phrases à variable(s) lexicale(s), par recherche et sélection de la phrase dans un sous-domaine de situation, puis choix de valeurs pour le ou les champs variables.

Un prototype SurviTra-1 a ensuite été développé (cf. figure 2.9), à l'aide de la plate-forme *Krater* (détaillée en partie 4), qui privilégie l'approche monoposte. Il est maintenant opérationnel pour le mode phrases alignées, sur un premier domaine ciblé (Restaurant, en Inde et en France). Il est en cours de finalisation pour l'intégration des premiers composants UNL disponibles (la construction de ces ressources logicielles UNL étant complexe et longue), avec une couverture variable dans les deux langues prioritaires du projet (français, hindi).

On s'oriente maintenant vers une interface paramétrable en fonction de l'espace disponible (taille de l'écran, du *smartphone* à l'écran d'ordinateur fixe, en passant par le *netbook*), et des préférences de l'utilisateur.

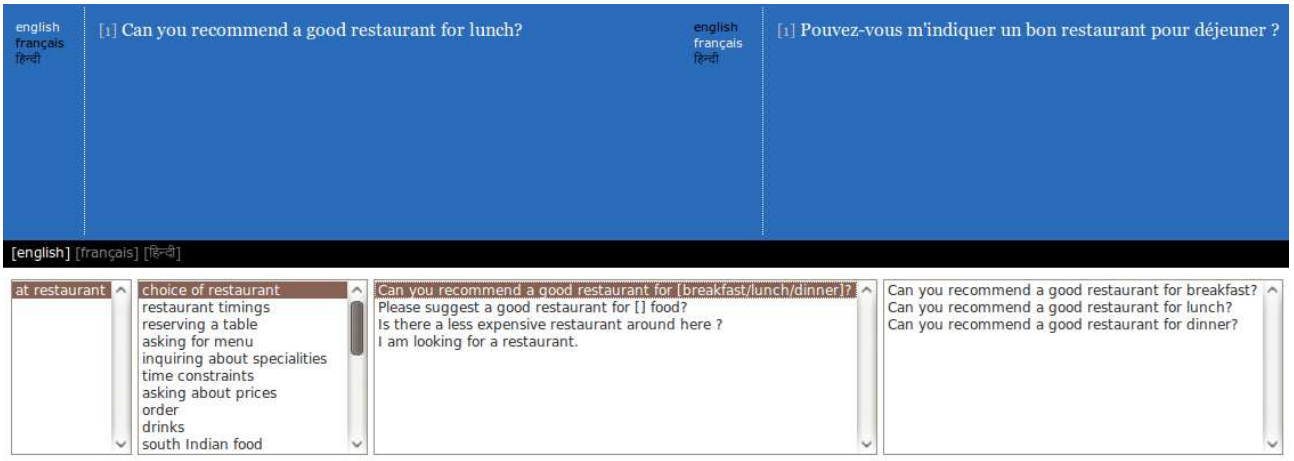


Figure 2.9 – Version 1 de l'interface de SurviTra.

L'interface de recherche des phrases est inspirée du *Finder* de *MacOS*, avec une colonne pour chaque niveau de l'arborescence du livre de phrases. Le dernier niveau est celui de l'instanciation des variables lexicales. Il s'agit d'une interface Web prévue pour PC « classique » (fixe ou portable). Dans une version PDA, chaque colonne de la recherche correspondrait à un écran. Une recherche « directe » par mot clef est aussi possible. Le champ de sélection est monolingue, mais la langue peut être modifiée à tout moment.

Les phrases sélectionnées et instanciées sont affichées dans un historique, en deux langues (elles aussi modifiables à tout moment).

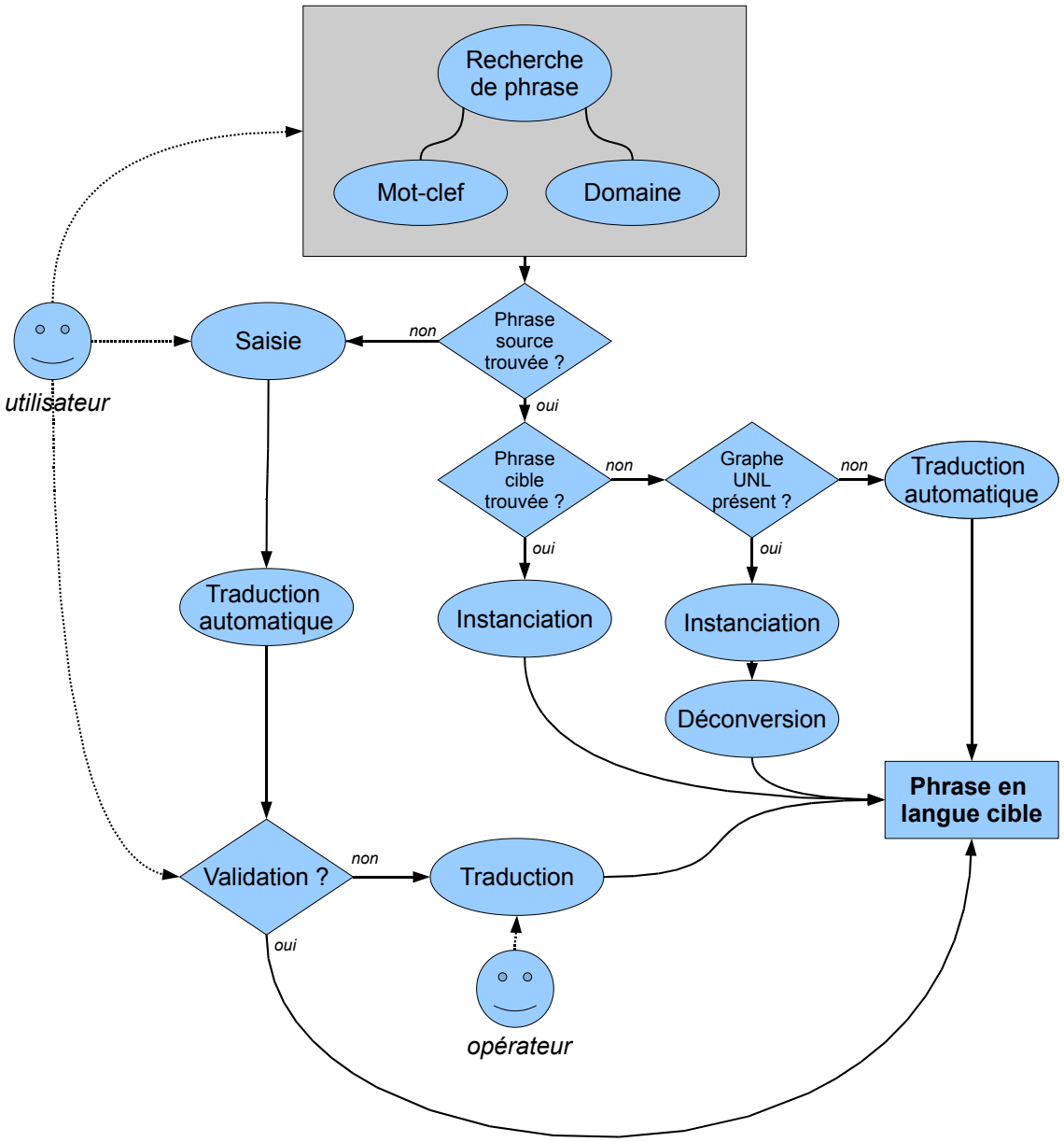


Figure 2.10 – Utilisation de SurviTra.

En vue d'une utilisation par les développeurs UNL, SurviTra intègre un « éditeur de corpus » léger (cf. figure 2.11), qui permet de modifier le livre de phrases (ajout, suppression, modification des phrases, lexies, sous-domaines, classes lexicales, et de leur structuration), pour un travail contributif « au vol », sans historique des mises à jour ni des intervenants, et en mode « bac à sable » : les modifications apportées sont aussitôt répercutées dans l'interface, mais uniquement pour la session en cours de l'utilisateur. Il est possible de télécharger ou d'importer une sauvegarde des modifications, dans un format CSV compatible avec les tableurs, ce qui permet aussi aux utilisateurs avancés une édition à l'aide d'un logiciel tiers.

Pour des modifications orientées vers les « campagnes de saisie / post-édition », et un travail collaboratif, un environnement convivial, a été développé par Cong Phap Huynh [CPBB08] : Sectra_W (« Système d'Exploitation de Corpus de Traductions sur le Web »). Il intègre des fonctionnalités wiki : suivi de version et gestion des modifications concurrentes. Les modalités de conversion de format entre les deux plateformes sont en cours de modélisation.

Dans le cadre de cette utilisation par les développeurs UNL, l'environnement SurviTra permet d'activer/d'inhiber les enconversions et déconversions disponibles, en *plugin*. Il fonctionne en banc d'essai-test-réglage de ces composants, tout ce qui est produit étant enregistré pour analyse et validation. Ce choix fonctionnel répond à la complexité de réalisation et de mise au point des déconversions et des enconversions : dans la réalité des réalisations de TA via UNL, celles-ci ne sont pas disponibles en même temps ni avec une même couverture syntaxique pour chaque langue. La plateforme fonctionne soit avec la mémoire de traduction (utilisant la base de phrases multilingue, dotée de graphes UNL, soit par traitement effectif UNL, selon les composants disponibles.

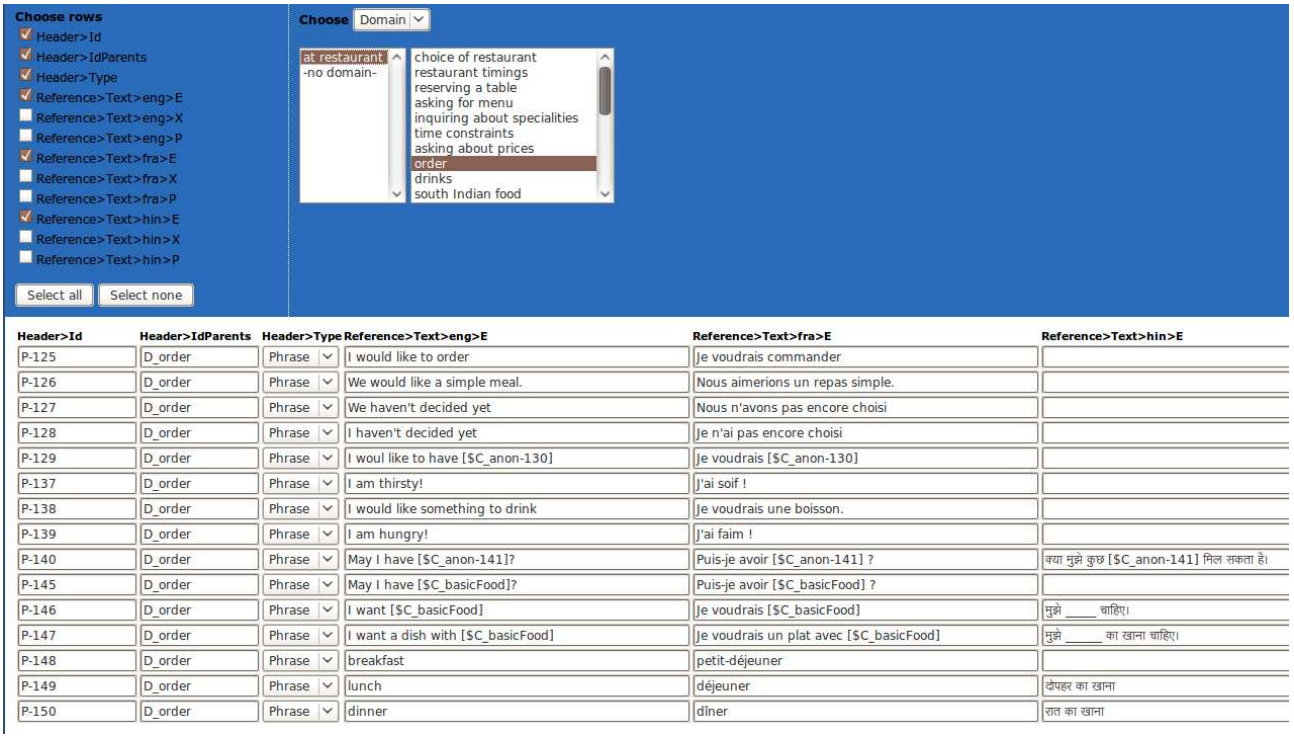


Figure 2.11 – Éditeur intégré à l'interface de SurviTra.

3.2 Prototypes de TA de dialogues oraux finalisés

3.2.1 C-STAR II

C-STAR[9] est un consortium international, créé en 1989 à l'initiative d'ATR[10], et réunissant plusieurs organismes (ATR, l'université Carnegie Mellon, Siemens, et l'université de Karlsruhe), dans l'objectif de réaliser des démonstrateurs de traduction de conversations visiophoniques dans le domaine touristique. En janvier 1993, une première démonstration eut lieu, et devant le succès rencontré, il fut décidé d'impliquer de nouveaux groupes de recherche dans le consortium, rebaptisé C-STAR II. C'est ainsi qu'en 1996, le CLIPS rejoignit le consortium, à la suite de l'institut coréen ETRI[11] (1993), et de l'IRST[12] italien (1994). D'autres groupes, affiliés et non membres, se sont associés à ces recherches, mais sans réaliser de démonstrateurs.

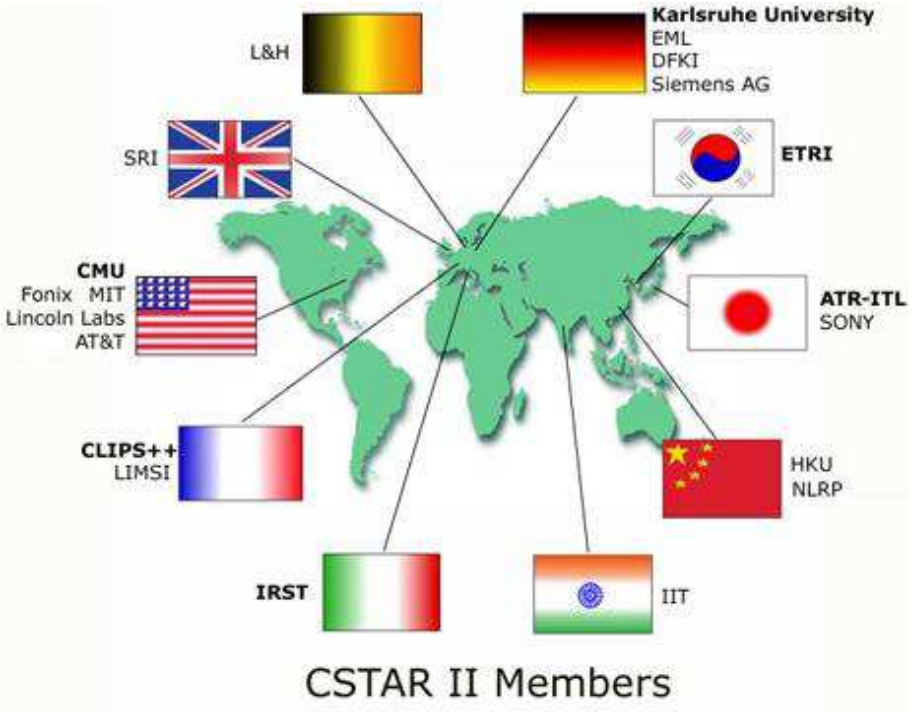


Figure 2.12 – Membres du projet C-STAR II (les membres ayant réalisé un démonstrateur sont en gras).

Les démonstrateurs réalisés au terme de C-STAR II (1999) peuvent traiter 6 langues : l'allemand, l'anglais, le coréen, le français, l'italien et le japonais. Le scénario d'utilisation est le suivant : un client, locuteur d'une langue donnée, appelle un agent de voyage, locuteur d'une langue différente ; les propos du client sont analysés

[9] *Consortium for Speech Translation Advanced Research.*

[10] *Advanced Telecommunication Research*, organisme de recherche semi-public japonais

[11] *Electronics and Telecommunication Research Institute*, organisme de recherche semi-public coréen.

[12] *Istituto per la Ricerca Scientifica e Technologica.*

et traduits vers la langue de l'agent, puis portés à sa connaissance via un synthétiseur vocal ; et inversement lorsque l'agent répond au client. De plus, chaque locuteur peut consulter une rétrotraduction écrite de ce qu'il vient de dire, lui permettant ainsi de contrôler si le système a bien « compris » ce qu'il a dit.

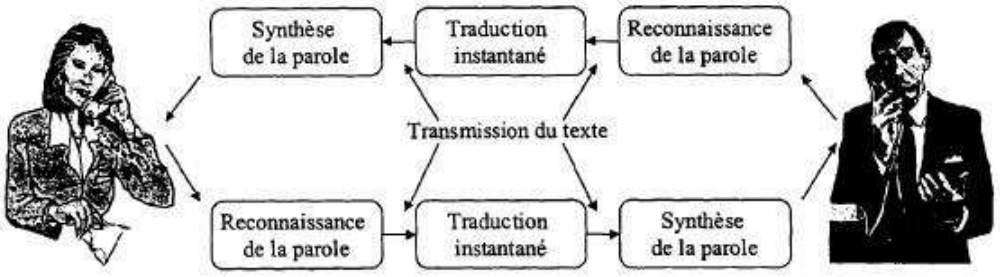


Figure 2.13 – Principe du système de TA de parole développé dans le cadre de C-STAR II [AJ98].

Modules

Chaque groupe de recherche développant son propre démonstrateur, leur fonctionnement peut diverger en fonction des choix techniques de chaque équipe. On s'intéressera ici principalement au démonstrateur français, développé par le CLIPS avec l'aide du LATL[13] (Genève) et du LAIP[14] (Lausanne).

Le module de reconnaisance de la parole, RAPHAEL [AJ98], a été développé au CLIPS par l'équipe GEOD, à partir de la boîte à outils Janus III [LWL97] réalisée par l'université Carnegie Mellon (CMU, Pittsburgh) et l'université de Karlsruhe. Il s'agit d'un système de reconnaissance multilocuteur à grand vocabulaire[15], à base de chaînes de Markov cachées, et qui repose sur une architecture client-serveur, le processus de reconnaissance proprement dit, gourmand en ressources, prenant place sur le serveur. Il produit une transcription orthographique, dépourvue de ponctuation et de marques d'accord (genre et nombre) lorsque ceux-ci ne sont pas marqués phonétiquement.

Le texte obtenu est traduit vers un langage formel pivot, l'IF (Interchange Format). Il s'agit d'un format de représentation sémantique, qui permet de représenter la signification d'un énoncé indépendamment de la langue dans laquelle il est formulé. Ce nouveau texte est ensuite traduit dans la langue cible.

Ce principe est utilisé par les démonstrateurs allemand, anglais, coréen, français et italien, ce qui permet de réduire le coût de développement des modules de traduction : dans un système de traduction par pivot, il suffit d'un seul module de

[13] Laboratoire d'Analyse et de Technologie du Langage.
[14] Laboratoire d'Analyse Informatique de la Parole.
[15] 10 000 mots (vocabulaire touristique) dans le cas de C-STAR.

traduction langue source/langue pivot par langue source, et d'un seul module de traduction langue pivot/langue cible par langue cible.

Le module français/IF a été réalisé par l'équipe GETA du CLIPS [BBC99] [BG00], sous Ariane-G5, un générateur de systèmes de traduction développé au GETA.

Le module IF/français a quant à lui été réalisé au LATL de Genève [WW98]. A l'aide d'une grammaire générative, il construit un texte en langue cible à partir de l'IF.

Enfin, le module de synthèse vocale a été développé au LAIP de Lausanne [KW97] [KZ98]. La conversion texte/phonèmes se fait par règles, et le signal audio est produit par le synthétiseur MBROLA de l'université de Mons (Belgique) [DL93] [DP96] [Dut96].

Au total, le système comporte cinq modules [Bla04a] :

un module de reconnaissance vocale qui prend en entrée le signal vocal produit par le locuteur, et fournit en sortie une chaîne de mots orthographique,

un module de traduction de cette chaîne de mots en une structure sémantique pivot (IF) commune à toutes les langues,

un module de communication qui permet de déposer cette IF dans une structure de données en vue de la génération dans d'autres langues et de lire une IF pour la traduire,

un module de génération qui prend en entrée l'IF et produit en sortie une chaîne de mots orthographique,

un module de synthèse qui convertit cette chaîne de mots, en un signal vocal.

Le format d'échange (IF)

Le module de traduction du français vers l'IF a été réalisé par l'équipe GETA [BBC99] [BG00] sous Ariane-G5, un environnement de programmation destiné aux créateurs de modèles de traduction automatique. Il prend en entrée une transcription orthographique d'un énoncé oral, et opère une traduction complète de cette transcription du français vers l'IF, avec des étapes classiques d'analyse du texte source (adaptée aux spécificités des entrées, qui divergent de l'écrit « bien formé » traditionnel), de transfert (adaptation à la langue cible des données recueillies) et de génération en langue cible.

Lors de la démonstration du jeudi 22 juillet 1999, ce module fonctionnait au centre de calcul d'IBM à Montpellier, sur un IBM 9672-RX5 à 60 Mips. Le temps de réponse, y compris les transferts par Internet (TCP/IP) et les manipulations de fichiers (codage/décodage) était de 5 à 7 secondes.

Interface
L'interface du système peut se décomposer en 11 zones [Bla04a] (cf. figure 2.14) :

1. le bouton PTT, « push to talk », lance le processus de reconnaissance ;

2. le bouton Envoi Hypo sert à valider une hypothèse de reconnaissance (présentée en 4) ;

3. le bouton Ignore sert à indiquer à l'interlocuteur qu'il doit ignorer la traduction reçue ;

4. hypothèse de reconnaissance, mise à jour au fur et à mesure que l'utilisateur parle ;

5. IF produites ;

6. rôle (client ou bien agent) ;

7. rétrogénération en français des IF produites localement ;

8. génération en français des IF produites pour celui qui répond (messages précédents) ;

9. langues affichées dans la trace multilingue (11) ;

10. dernières IF reçues ;

11. générations produites par les sites distants.

Démonstration
Le démonstrateur français [BBC99] a été présenté au public le jeudi 22 juillet 1999 [Bla04a], dans les locaux de l'IMAG à Grenoble, en interaction avec les systèmes allemand, américain et coréen (recourant au même système de traduction par pivot). Le scénario mis en œuvre est simple : un client souhaite organiser un voyage en Allemagne, aux États-Unis, puis en Corée du Sud ; il réserve des vols, des chambres, s'informe sur les tarifs, les conditions météo, les activités possibles, etc. auprès d'agents de voyage de chaque pays. De nombreux représentants de l'industrie et des centres de recherche étaient présents (IBM, Xerox, Digigram, Azimut, France Telecom, Magellan Ingénierie, le CNET, Spacio Guide), ainsi que de l'ANVAR. La couverture médiatique fut importante, aussi bien dans la presse écrite (le Monde, Libération, Science & Vie, etc.), qu'à la radio (France Info, France Inter, RMC, etc.) et à la télévision (France 3 régional, M6).

Perspectives
A l'issue de C-STAR II, les partenaires européens et américains du consortium souhaitaient aller plus loin, et c'est ainsi qu'en 2000 le projet Nespole ! fut lancé conjointement par la Commission Européenne et la *National Science Foundation Division of Mathematical Sciences* américaine. Sur le plan technique, ce projet marque un progrès par rapport à C-STAR II par l'utilisation de protocoles et de formats standard. Il a aussi donné lieu à des travaux d'évaluation plus poussés.

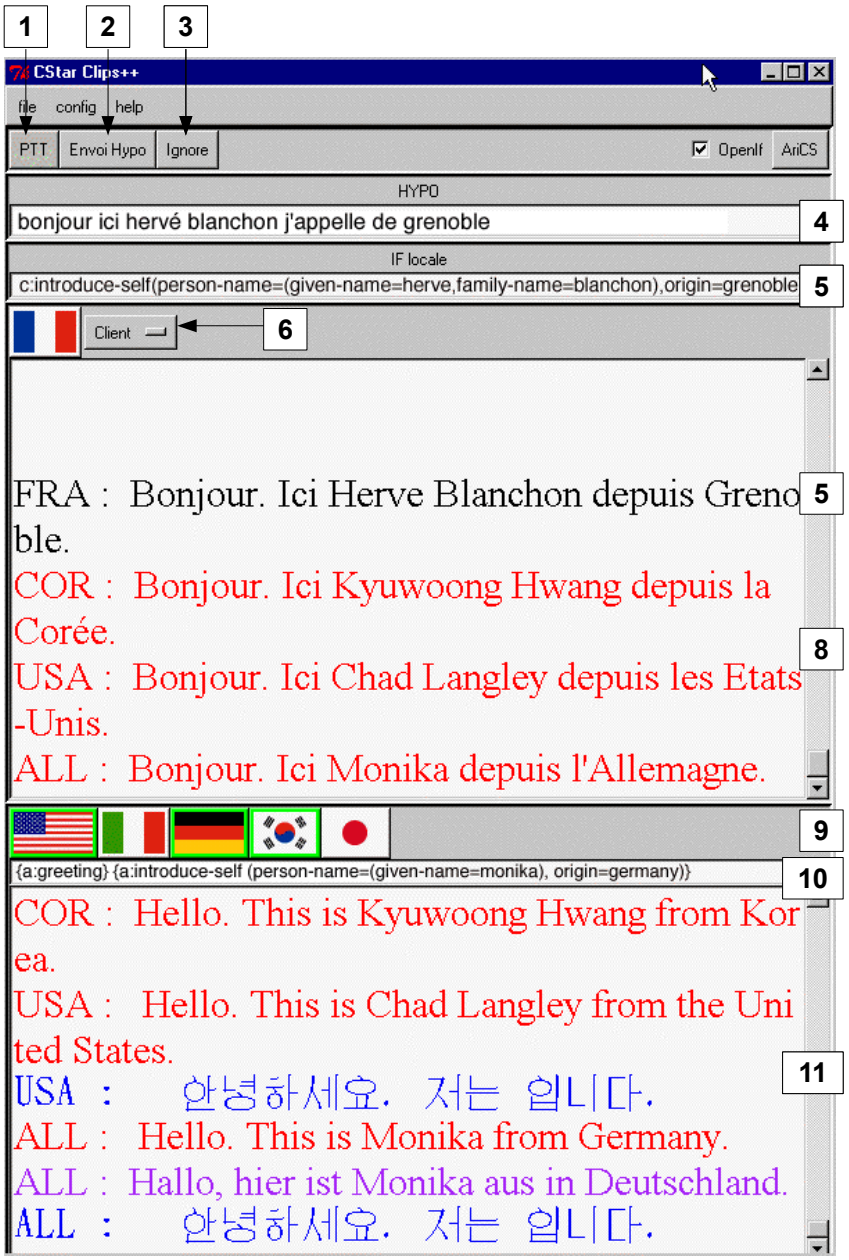


Figure 2.14 – Interface du démonstrateur français de C-STAR II, ici lors d'une conversation à 4.

3.2.2 Nespole !

Le projet Nespole ! (*NEgociating through SPOken Language in E-commerce*) [MMS+02] a vu le jour à l'issue de C-STAR II, certains des partenaires de ce projet souhaitant aller plus loin. En effet, dans C-STAR II, l'attention était focalisée sur les aspects TALN du projet, au détriment de l'IHM et de l'architecture globale du système [Bla04a]. L'interface obtenue, quoique fonctionnelle, était lourde et peu pratique, et le système s'avérait complexe à mettre en place pour les utilisateurs comme

pour le prestataires de services de TA.

Dans le cadre du projet Nespole !, les partenaires ont donc tenté de mettre en œuvre les acquis de C-STAR II dans un contexte d'application commerciale plus concret et utilisable. Ce projet, soutenu par la Commission européenne et la *National Science Foundation Division of Mathematical Sciences* américaine, rassemblait quatre laboratoires de recherche : l'IRST (Trente, Italie), l'ISL (Karlsruhe, Allemagne), le CLIPS, et l'université Carnegie-Mellon (Pittsburgh, Etats-Unis) ; ainsi que deux partenaires industriels : l'office du tourisme de Trente, et Aethra, une société italienne de télécommunications.

L'objectif était de réaliser un système robuste et multimodal pour la traduction automatique en e-commerce, en allemand, anglais, français et italien, dans le cadre d'un scénario bien précis : un touriste étranger (germanophone, anglophone ou francophone), souhaitant se rendre dans la région de Trente (Italie) pour des vacances d'hiver, et contactant l'office du tourisme afin de préparer son séjour.

Mise en œuvre
Le système client fonctionne sur tout PC/Windows équipé d'un microphone, de haut-parleurs, d'une connexion à Internet à 64 kbps (mais nettement moins sans la vidéo), et d'un logiciel de VoIP, tel que NetMeeting de Microsoft. La reconnaissance vocale et la traduction étant effectuées sur un serveur, aucune puissance de calcul importante n'est requise sur le poste client. Du côté serveur, le processus le plus lourd, la reconnaissance vocale, s'exécutait lors de la démonstration sur un Pentium III à 1 GHz, et s'effectuait en temps réel. En pratique, et sauf congestion du réseau, les traductions arrivaient moins d'une seconde après que l'utilisateur avait fini de parler [MMS+02]. Pour le français, on avait donc progressé d'un facteur 5 par rapport à C-STAR II.

Architecture
L'élément central du réseau mis en place pour la visioconférence (cf. figure 2.15) est le médiateur. C'est ce module qui sert d'intermédiaire entre les clients de visioconférence et les serveurs de reconnaissance et de traduction, ce qui simplifie la procédure de connexion pour le client. Les connexions médiateur – serveurs de traduction se font par *sockets* (indépendants de la plate-forme) et, comme on peut le constater, les protocoles de communication correspondent au standard de la voix sur IP, le H323, qui transmet le son suivant le protocole G711 et l'image suivant le protocole H261, tous deux compressés.

Le format d'échange (IF)
L'IF de Nespole ! a été développé à partir de celui de C-STAR II, et comprend nettement plus d'éléments : 75 actes de parole, 123 concepts et 189 arguments. La spécification de cet IF est fondée sur les données, et a été étendue au fil du projet, à chaque fois que le besoin s'en faisait sentir.

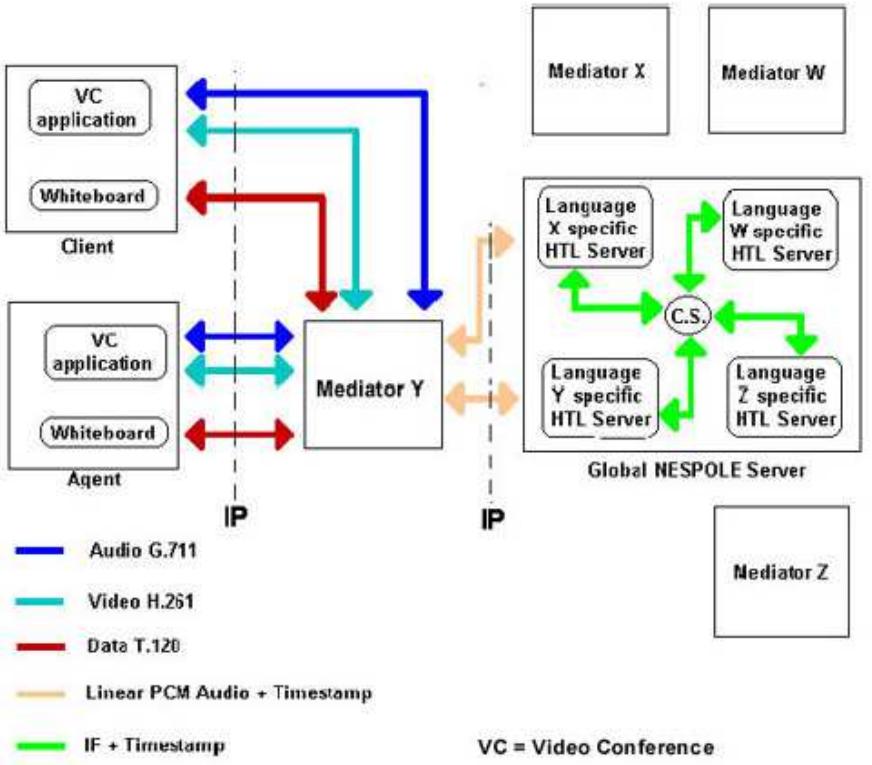


Figure 2.15 – Architecture informatique de Nespole !.

L'énoncé « je voudrais une chambre simple du 10 au 15 septembre » peut être représenté en IF sous la forme suivante :

```
{ c:give-information+disposition+room
     ( disposition=(desire, who=i),
       room-spec=(identifiability=no, single\_room),
       time=
       ( start-time=(md=10),
         end-time(md=15, month=9)
       )
     )
}
```

On peut décomposer cet énoncé en 3 parties :

1. « je voudrais »

   acte de parole `give-information`

   concept `+disposition`

   argument `disposition=(desire, who=i)`

2. « une chambre simple »

concept `+room`

argument `room-spec=(identifiability=no, single_room)`

3. « du 10 au 15 septembre »

   argument `time=(start-time=(md=10), end-time(md=15, month=9))`

Interface utilisateur

L'interface utilisateur de Nespole ! (cf. figure 2.16) peut se décomposer en 6 zones principales [Bla04a] :

1. environnement de vidéoconférence (ici NetMeeting) ;
2. *system hears* : retour de la reconnaissance vocale en langue source, permettant de contrôler la qualité de cette dernière, et le cas échéant de signaler au système une erreur et de recommencer (*Cancel Translation*). Sinon, il clique sur *Send* pour lancer la traduction de l'énoncé reconnu.
3. *system understands* : rétrotraduction, permettant à l'utilisateur de vérifier la qualité de la traduction. Le bouton *Cancel* Translation est aussi utilisable pour signaler que la traduction est incorrecte et que l'on va recommencer ;
4. contrôle de l'ouverture et de la fermeture du microphone ;
5. tableau blanc, offrant de nombreuses possibilités d'interaction, et en particulier l'échange et l'annotation en temps réel par les utilisateurs de documents graphiques tels que des cartes, des plans, etc.
6. historique du dialogue (SH=*system hears*, SU=*system understands*, ST=*system translates*).

Principaux modules

Le module de reconnaissance utilisé est le même que pour C-STAR II, à savoir RAPHAEL. Il a été adapté à la perte d'information en entrée due à la compression des données audio, et conçu de sorte à pouvoir être facilement étendu.

Le module d'analyse du français recourt à une version augmentée de l'IF de C-STAR II, mais le procédé d'analyse, basé sur des patrons, est différent : il s'agit de faire une analyse par îlots du résultat de la reconnaissance de parole. Cette méthodologie se révèle bien adaptée à la langue parlée, moins continue que la langue écrite d'un point de vue syntaxique [Bla04a]. Par rapport aux démonstrateurs des autres équipes participant au projet, l'analyseur du CLIPS se distingue par son approche linguistique (par expressions régulières), imposée par le choix de l'environnement ARIANE-G5, alors que les autres équipes ont privilégié une approche stochastique (probabiliste) [BBF+01]. Le vocabulaire de l'analyseur couvre 2028 lemmes.

Contrairement à ce qui c'était passé pour C-STAR II, c'est le CLIPS qui était en charge de la génération pour Nespole !. Le générateur [Bla04a] utilise un ensemble de phrases à trous associées à un acte de dialogue.

Figure 2.16 – Interface utilisateur de Nespole !, lors d'un dialogue anglo-italien.

Pour la synthèse vocale, le texte orthographique est converti en phonèmes par le système EULER de l'université polytechnique de Mons (Belgique) [BDM+00], puis passé à MBROLA (déjà utilisé pour C-STAR II) pour la synthèse. Ces outils ont été déployés par le CLIPS.

Évaluation

Une évaluation des différents démonstrateurs a aussi été mise en œuvre, avec et sans reconnaissance vocale (pour ne tester que les performances du module de traduction). Cette évaluation portait sur deux dialogues extraits de la seconde collecte Nespole ! [MBC+03], et les juges choisis pour évaluer l'acceptabilité (la compréhensibilité) des énoncés étaient des étudiants en dernière année d'école de traduction (niveau master 2). Les tableaux 2.4 à 2.7 présentent les résultats obtenus. Les locuteurs français, anglais et allemands étaient des clients, et les locuteurs italiens des agents.

	Français	Anglais	Allemand	Italien
WAR	58%	56%	51%	76%
Paraphrase acceptable	60%	67%	62%	76%

Table 2.4 – Évaluation de Nespole ! : reconnaissance automatique de parole.

	Français	Anglais	Allemand	Italien
Transcription manuelle	77%	68%	61%	51%
Reco. automatique de parole	58%	50%	51%	42%

Table 2.5 – Évaluation de Nespole ! : rétrotraduction.

	Français-Italien	Anglais-Italien
Transcription manuelle	77%	70%
Reco. automatique de parole	58%	50%

Table 2.6 – Évaluation de Nespole ! : traduction client vers agent. Les chiffres pour la traduction allemand-italien ne sont pas disponibles à cause d'un comportement inconsistant des juges.

	Italien-Français	Italien-Anglais	Italien-Allemand
Transcription manuelle	37%	33%	45%
Reco. automatique de parole	33%	30%	38%

Table 2.7 – Évaluation de Nespole ! : traduction agent vers client.

3.2.3 Verbmobil

Verbmobil [Wah00] (1993-2000) est un projet allemand financé par le ministère de l'Éducation, des Sciences, de la Recherche et de la Technologie, dont l'objectif était le développement d'un système mobile pour la traduction automatique de parole, en allemand, anglais et japonais, lors de dialogues commerciaux : organisation de voyages, prises de rendez-vous, maintenance de PC à distance.

Réunissant plus d'une trentaine de groupes de recherches répartis entre l'Allemagne, les États-Unis et le Japon, et de grandes firmes industrielles (Alcaltel, Daimler-Benz, Debis Systemhaus, IBM, Philips, Siemens), sous la direction du Centre de recherche allemand pour l'intelligence artificielle[16], ce projet a reçu une aide financière totale de 60 millions d'euros sur 8 ans (1993-2000) [Bla04a].

Scénario

Il s'agit de conversations sur téléphone cellulaire (son à 16 kHz), avec traduction bidirectionnelle entre l'allemand, l'anglais et le japonais, portant sur les prises de rendez-vous (vocabulaire jusqu'à 6 000 mots), la planification de voyages (vocabulaire jusqu'à 1 000 mots) et la maintenance informatique (vocabulaire jusqu'à 30 000 mots).

Architecture

Le projet Verbmobil implique de nombreux modules de traitement, conçus par des

[16] *Deutsches Forschungszntrum für Künstlische Intelligenz*, Saarbruck.

équipes différentes dans des langages différents, aussi variés que C, C++, LISP, Prolog et Tcl/Tk [KNK00]. Globalement, chaque module correspondait à une équipe de développement.

Une première maquette s'appuyait sur une architecture multi-agents pour faire coopérer ces différents modules. Les agents communiquaient en s'envoyant des messages via un composant « boîtes aux lettres » central. Des librairies développées pour les différents langages permettaient de standardiser les échanges. Cette approche révéla assez rapidement des défauts rédhibitoires :

- les modules devaient connaître l'ensemble des composants et maintenir la liste des modules produisant les données qui leur étaient nécessaires ;
- manque de flexibilité, tout ajout de module entraînant une modification de tous ceux qui devaient interagir avec lui ;
- complexité due au grand nombre de canaux.

C'est pourquoi une seconde architecture fut mise en place. L'organisation en agents fut abandonnée, et la coopération entre modules, qui ne pouvaient plus communiquer directement, s'effectue par le biais d'un tableau noir, et s'organise en trois parties :

1. un groupe de composants, experts dans un domaine (du même ordre que pour la première maquette) ;
2. un tableau noir (base de données partagée) dans lequel les modules peuvent lire et écrire ;
3. un contrôleur, qui décide de l'ordre dans lequel les modules sont exécutés.

Le contrôleur décide dynamiquement de l'ordre d'exécution des modules, en fonction des besoins et des capacités exprimés par ces derniers. Le résultat de la reconnaissance de parole est soumis à 3 types d'analyseurs différents. Ces analyseurs produisent chacun une structure qui est souvent partielle (en fonction de ce qui a pu être analysé), et ces structures sont combinées pour obtenir une représentation sémantique complète. Le résultat est transmis à 5 systèmes de traduction, correspondant à autant de méthodologies, et qui associent un degré de confiance à leurs traductions. La meilleure traduction est sélectionnée en fonction de ce degré de confiance, et transmise au module de synthèse.

Évaluation

L'évaluation a été menée sur des scénarios de planification de voyage, avec un vocabulaire de 10 000 mots. Au cours de l'évaluation, dont les résultats sont présentés dans le tableau 2.8, il a été constaté une divergence intéressante entre ce que les utilisateurs et les linguistes considéraient comme une traduction acceptable : en effet, les traductions comportant moins d'informations que l'original étaient souvent jugées acceptables, contrairement aux traductions comportant plus d'information, généralement jugées mauvaises.

WAR		*>50%*	*>75%*	*>80%*
Allemand → Anglais	Nombre de tours de parole concernés	5069	3267	2723
	Sélection automatique de la meilleure traduction	57%	66%	68%
	Sélection manuelle de la meilleure traduction	88%	95%	97%
Anglais → Allemand	Nombre de tours de parole concernés	4136	3254	2291
	Sélection automatique de la meilleure traduction	53%	58%	60%
	Sélection manuelle de la meilleure traduction	86%	92%	94%

Table 2.8 – Pourcentage d'énoncés acceptables, en fonction du taux de reconnaissance lexicale (*WAR*), dans le cas de choix automatique et manuel de la meilleure traduction parmi les 5 proposées par les systèmes de traduction

3.2.4 Études en cours sur l'outillage et la collecte de dialogues bilingues (projet ERIM)

À la suite d'une expérimentation longue conduite sur une plate-forme d'ATR[17], c'est avec le chinois qu'a débuté, au CLIPS, l'étude et l'élaboration d'outils génériques d'aide à la communication orale multilingue multimodale à distance, en collaboration avec les laboratoires NLPR-IA de l'Académie des Sciences de Chine et LIAMA, dans le cadre des projets ERIM [Faf04b] [Faf04a] [FBSZ04] et ChinFaDial.

Des prototypes successifs d'ERIM-Interprète et ERIM-Collecte ont été développés. Il s'agit de plates-formes pour l'interprétariat (humain) sur réseau, avec collecte, de corpus de dialogues spontanés bilingues traduits, à architecture distribuée : 2 stations "locuteur" (ou plus), 1 station "interprète" (ou plus), 1 serveur de communication et 1 serveur de collecte de corpus.

ERIM[18] [Faf04b] [Faf04a] [FBSZ04] est une plate-forme d'expérimentation, destinée à tester différents aspects de la communication bilingue sur le Web, développée au cours de divers projets, notamment en partenariat avec le laboratoire franco-chinois LIAMA[19] et le NLPR[20] (tous deux à Pékin). Cette plate-forme est actuellement constituée de quatre composants, correspondant à plusieurs types d'applications possibles des systèmes de parole bilingue sur le Web.

Un trentaine d'heures de dialogues spontanés bilingues traduits, sur domaine finalisé (réservation hôtelière), a été produite avec ERIM (en français-chinois, français-

[17] Advanced Telecommunications Research, Kyoto, Japon
[18] Environnement Réseau pour l'Interprétariat Multimodal
[19] Laboratory In computer science, Automation and applied MAthematics
[20] National Laboratory of Pattern Recognition

vietnamien, français-hindi et français-tamoul) sur le terrain (Chine, Vietnam, Inde), et à Grenoble.

ERIM-Interprète

ERIM-Interprète (cf. figures 2.17 et 2.18) sert de base aux autres composants. Il s'agit d'une application client-serveur, développée en Tcl/Tk, qui permet d'explorer des scénarios assez différents de ceux de l'interprétariat classique ; notamment l'« interprétariat intermittent à la demande » : les locuteurs essaient de converser en utilisant la connaissance qu'ils ont de la langue de leur interlocuteur, ou dans une langue véhiculaire, ou connue des deux ; lorsque cette communication s'avère impraticable (maîtrise de la langue, accent, souhait d'une traduction par un tiers), ou pour des séquences « sensibles » de leur échange, ils font appel momentanément aux services d'un interprète disponible sur le Web, qui peut les aider.

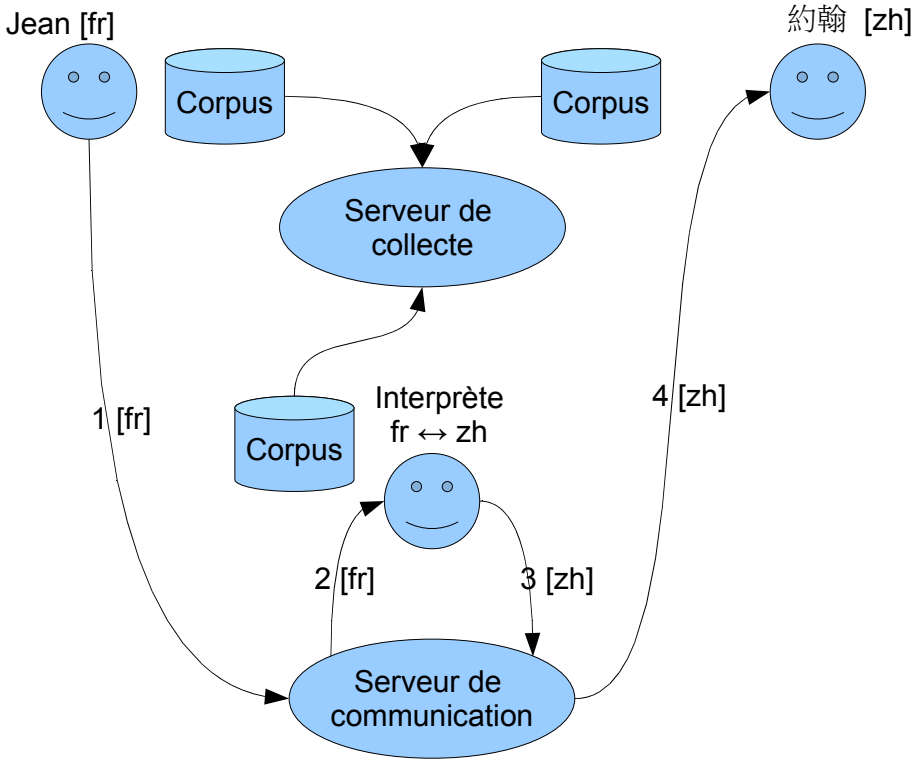


Figure 2.17 – Principe de fonctionnement d'ERIM interprète. Les locuteurs (ici francophone et sinophone) peuvent dialoguer verbalement (étapes 1 et 4). En cas de difficulté, ils peuvent demander de l'aide à un interprète (ajout des étapes 2 et 3). Chaque tour de parole, y compris les méta-données du dialogue, est enregistré sur le client, et sur un serveur de collecte.

ERIM-Collecte

ERIM-Collecte est destiné à la collecte de corpus de parole. Ce composant est basé sur ERIM- Interprète, auquel il adjoint une fonction d'enregistrement des dialogues. Ceux-ci sont collectés localement, puis transmis automatiquement au serveur de collecte.

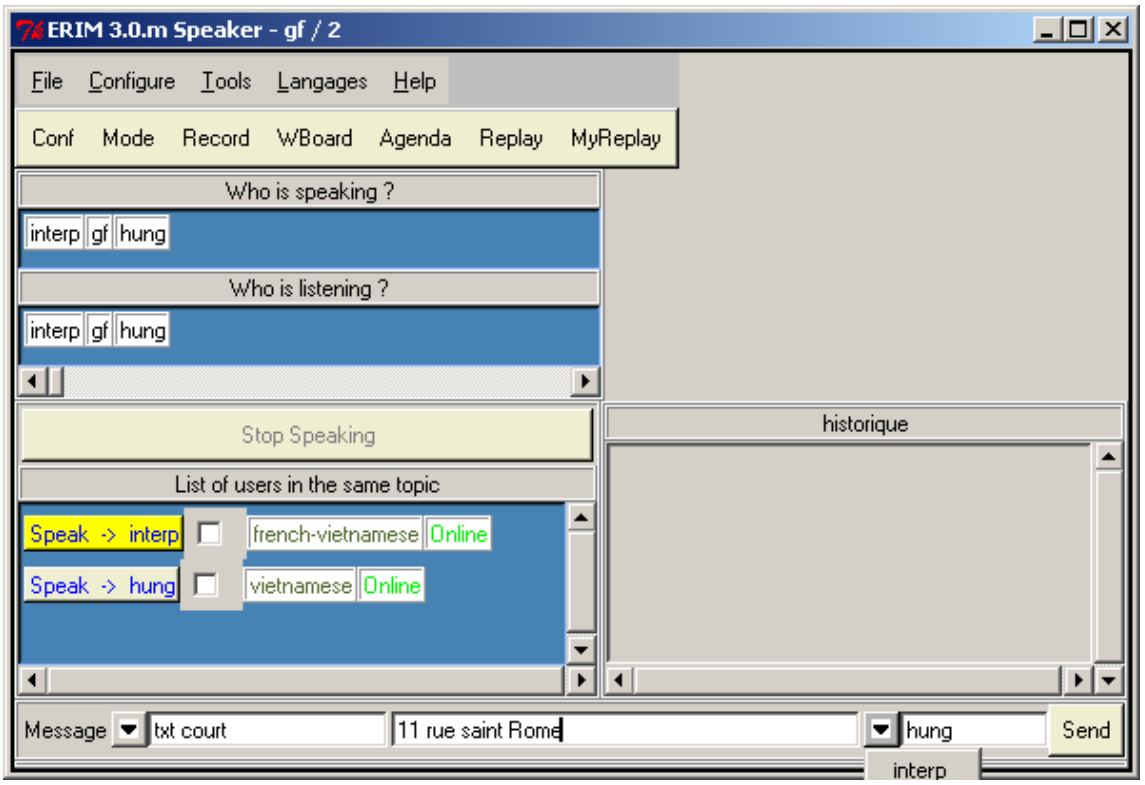


Figure 2.18 – Interface d'ERIM collecte-interprète (version 3.0)

ERIM-TA

L'objectif de ce composant est de permettre l'introduction de *plugins* de traduction automatique de la parole (reconnaissance, traduction, synthèse). Il a été développé en partenariat avec *Spoken Translation Inc.* (Berkeley, Etats-Unis). Ce composant est basé sur des systèmes commerciaux, de *Philips* (reconnaissance), *Linguatec* (traduction), et *ScanSoft* (synthèse). Une fonction de rétrotraduction permet à l'utilisateur d'évaluer et éventuellement de corriger l'analyse de ses énoncés.

Considére comme un environnement de test et de réglage de composants de TA, il utilise l'interprète humain, soit pour produire une version bilingue de référence, soit en *backup* (en «ange gardien») en cas d'expérimentation des modules de TA en situation réelle.

Cette version expérimente diverses aides à la TA de parole, aussi bien des « aides à la communication multimodale » (1 à 2), des « aides à la gestion des situations d'interprétariat en ligne » (3), que des aides « linguistiques » (4 à 6) :

1. visioconférence ;
2. partage de données (notamment via un tableau blanc) ;
3. planification et gestion de rendez-vous (sur un agenda) entre locuteurs et interprètes ;
4. recherche dans des dictionnaires bilingues ;

5. reconnaissance de la parole produisant une trace de la conversation ;
6. TA de parole, partielle ou complète.

La facette ERIM-Aides de l'environnement, orientée vers des aides en ligne à la communication multilingue sur réseau, est en début de prototypage. Les aides aux « interprètes en ligne » (professionnels ou occasionnels) peuvent être lexicales, terminologiques, de prononciation... Il peut s'agir également d'aides à l'interaction (multimodalité, utilisation d'un tableau blanc pour le partage entre locuteurs de documents visuels ou graphiques avec marquages à main levée, vision de l'interlocuteur par webcam en situation de dialogue à deux ou pour des applications de visioconférence multilingue, transcription instantanée pour rappel visuel du texte des derniers tours de parole...), ou d'aides à la gestion de l'activité de l'interprète, à la prise et à la consultation (sur le vif, ou après et avant une séquence de traduction orale) de notes personnelles, écrites ou orales...

Les aides aux locuteurs, prévues mais pas encore implémentées, sont comparables, avec à terme l'adjonction de composants de « Traduction partiellement automatique » de parole (traduction automatique « aidée par l'humain » [BB94]) facilitant une communication médiatisée (de la meilleure qualité possible) dans des situations où il faudrait passagèrement se débrouiller sans l'interprète.

3.3 Systèmes commerciaux de TA de dialogues oraux

3.3.1 MedSLT

MedSLT est un système de traduction libre [BRC+05] [SBC+05] [RBC+06]. Le système est restreint à un sous-domaine (maux de tête, douleurs à la poitrine et à l'abdomen) de dialogue finalisé médecin-patient, et cible en particulier les situations d'urgence. Il supporte l'arabe, l'anglais, le catalan, l'espagnol, le français et le japonais.

En pratique, le médecin doit d'abord s'entraîner à utiliser le système. En utilisation, il commence par sélectionner un sous-domaine et une paire de langues, puis peut poser des questions relevant de ce sous-domaine au patient. Une rétrotraduction lui est présentée, qu'il peut invalider, mais pas corriger ou désambiguïser. S'il n'invalide pas la rétrotraduction, une traduction est générée et synthétisée. Le patient se contente de répondre « oui » ou « non », ou de désigner une partie de son corps. Un mode bidirectionnel est envisagé [BFS+07], les réponses du patient pourront alors être traduites. Une version portable est aussi disponible [SBC+05] (figure 2.20).

Architecture

Le module de reconnaissance vocale est basé sur le reconnaisseur de mots Nuance, et supporte entre 350 et 1 000 mots, en fonction du sous-domaine [RBB+08]. Les treillis de mots sont ensuite traités de deux manières : à l'aide de grammaires, qui génèrent directement une représentation linguistique abstraite, et à l'aide d'un modèle de langage probabiliste, qui produit une phrase, à partir de laquelle la représentation

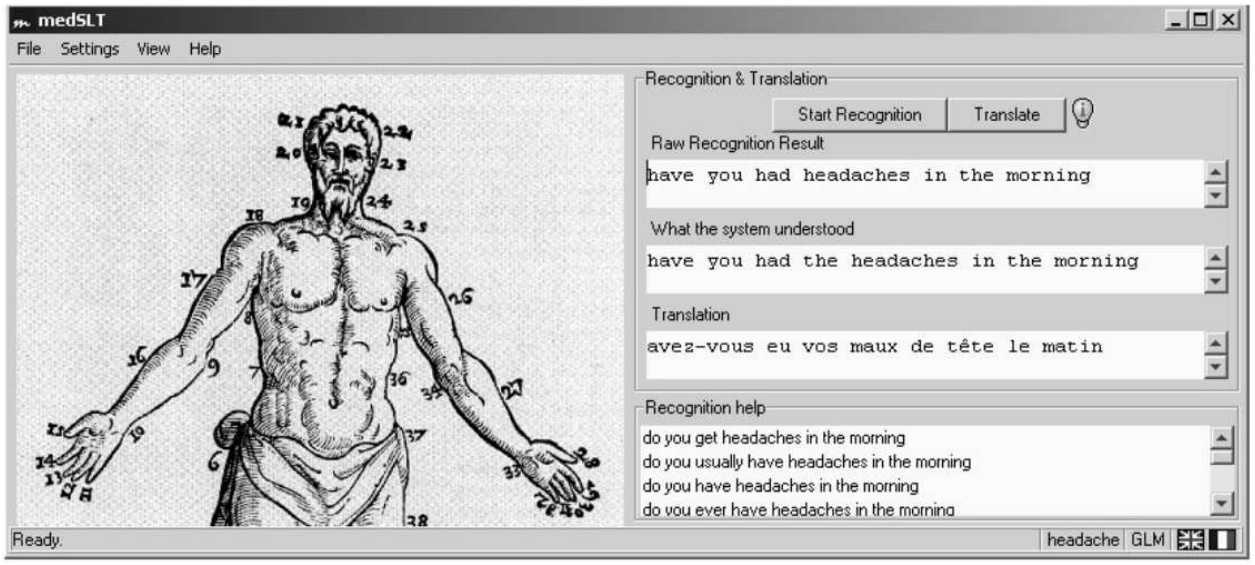


Figure 2.19 – Interface de MedSLT. Traduction de la phrase *do you ever have headaches in the morning ?* vers le français [BRC+05].

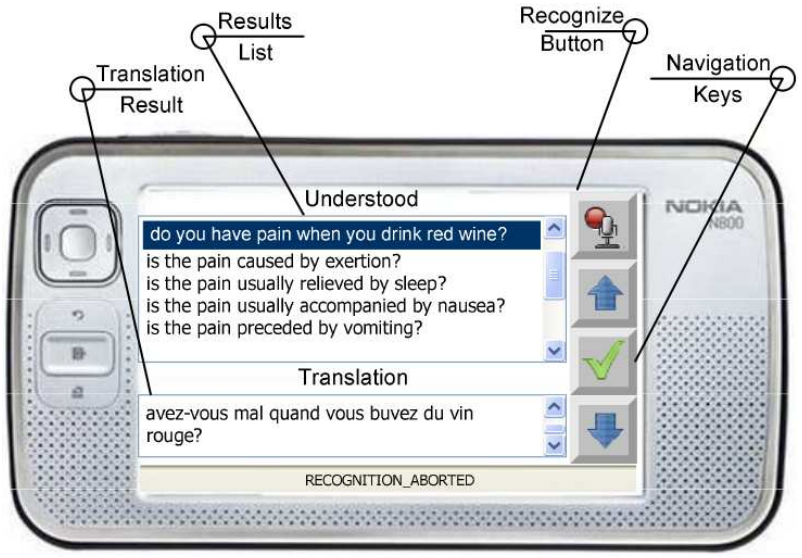


Figure 2.20 – Version portable [SBC+05].

linguistique abstraite sera générée. Cette représentation linguistique est de type «pivot hybride», c'est à dire qu'elle utilise des attributs et relations interlingues, et des unités lexicales de chacune des langues (cf. figure 2.21). Les meilleures hypothèses du système probabiliste sont présentées à l'utilisateur (icône « ı » dans l'interface), ce qui aide l'utilisateur à s'adapter au système de reconnaissance vocale.

Évaluation

Une évaluation a été menée en 2005 pour les couples anglais-français et anglais-japonais, dans le domaine des maux de tête [BRC+05]. Le système de reconnaissance a été entraîné sur un corpus de 575 questions standard fournies par un professionnel.

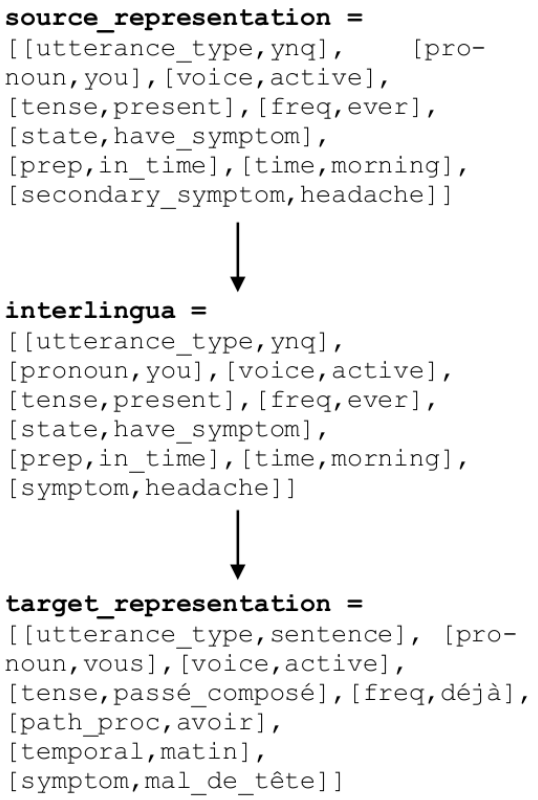


Figure 2.21 – Étapes de transfert. Traduction de la phrase *do you ever have headaches in the morning ?* vers le français [BRC+05] [RBB+08].

Les douze testeurs anglophones ont tout d'abord eu l'occasion de se familiariser avec le système, avant de jouer le rôle du médecin pour résoudre un problème simple : classer un mal de tête parmi huit types possibles. La moitié utilisaient le système de reconnaissance par règles, et les autres le système statistique. 870 tours de parole ont ainsi été collectés. Les résultats de cette évaluation sont présentés dans la figure 2.22 ; on constate que même sur un domaine très limité, le principal problème en utilisation réelle reste la reconnaissance vocale.

	French		Japanese	
	GLM	**SLM**	**GLM**	**SLM**
Bad Recognition	54.6%	59.8%	54.6%	59.8%
Good Translation	34.4%	30.8%	36.4%	32.8%
Acceptable Translation	8.7%	7.7%	3.6%	3.3%
Bad Translation	0.3%	0.2%	0.5%	0.5%
No Translation	2.0%	1.5%	4.9%	3.7%

Figure 2.22 – Évaluation de MedSLT pour le français et le japonais [BRC+05] (2005).

Une nouvelle évaluation a été présentée en août 2008, portant sur la reconnaissance vocale, pour l'anglais, le français et le japonais, toujours dans le domaine des maux de tête, et donne des résultats encourageants [RBB+08], présentés dans la table 2.9.

	Anglais	Français	Japonais
Vocabulaire (formes)	447	1025	422
Taux d'erreur lexicale (WER)	6%	8%	3%
Taux d'erreur sémantique	11%	10%	4%

Table 2.9 – Évaluation de la reconnaissance automatique de parole pour MedSLT [RBB+08] (2008).

3.3.2 Converser for Healthcare

Converser for Healthcare est un système commercial de traduction de parole édité par *Spoken Translation* [SZ05]. Il s'agit d'un système de dialogue «tout-venant» anglais-espagnol bidirectionnel, entre médecins anglophones et patients (et leurs familles) hispanophones, reposant sur des composants commerciaux (notamment Dragon Naturally Speaking pour la reconnaissance vocale). Il se distingue par la présence d'une phase de validation et, facultativement, de désambiguïsation interactive (l'utilisateur peut passer cette étape).

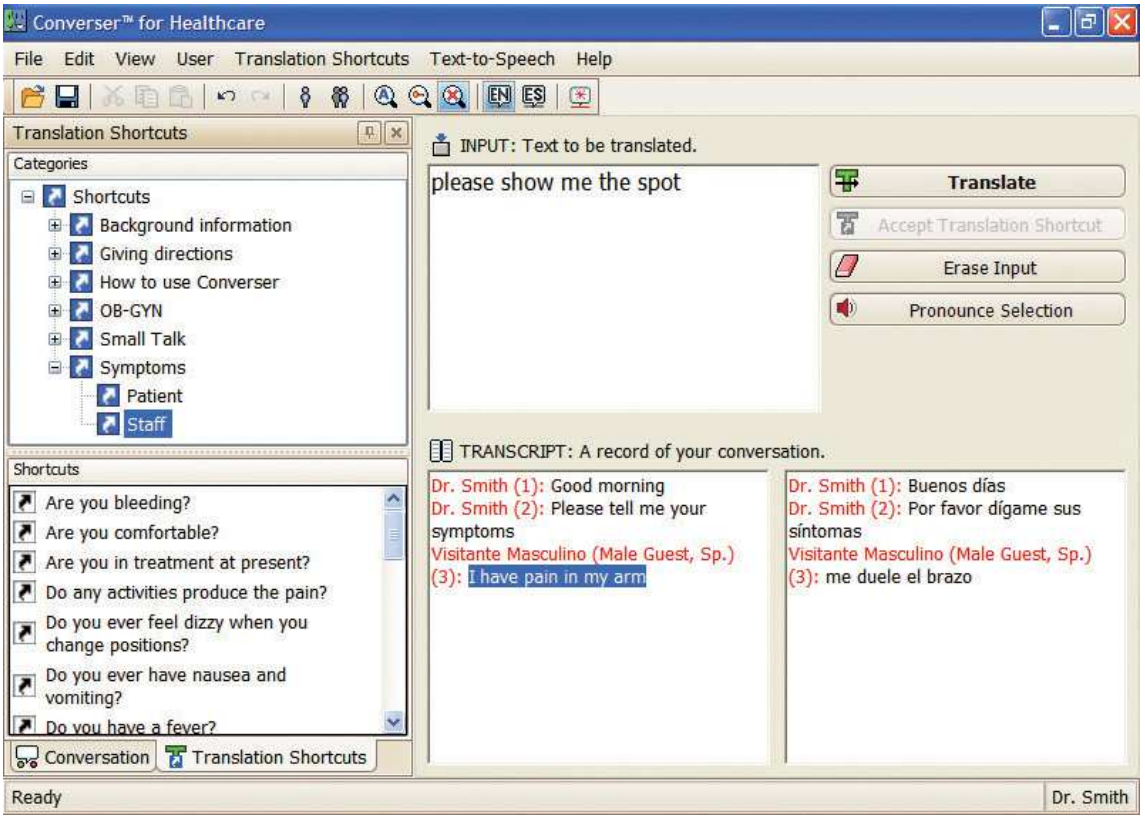


Figure 2.23 – Interface principale de Converser.

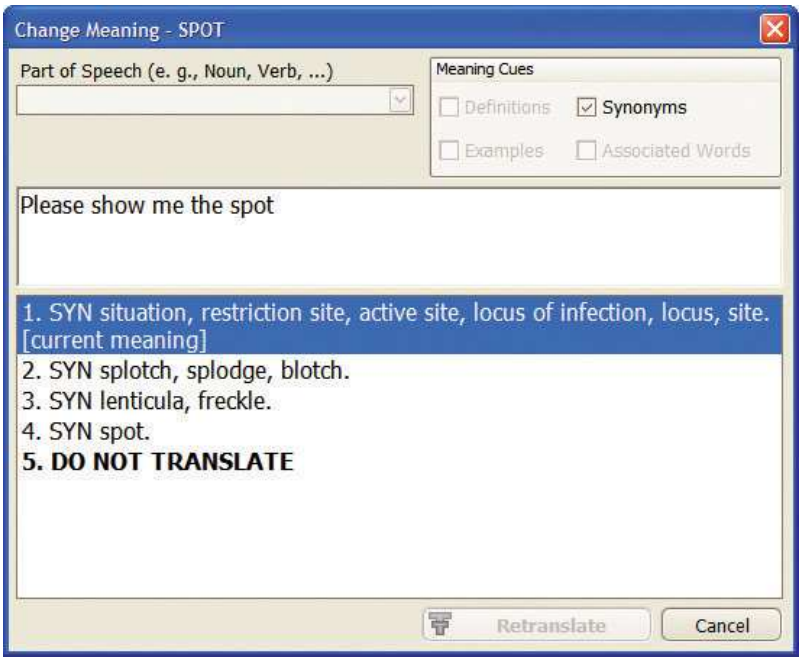


Figure 2.24 – Désambiguïsation interactive dans converser-4.

3.4 Les tchats multilingues

3.4.1 État de l'art

Il existe actuellement de nombreux outils de tchat multilingue, comme *WordLingo Chat*, *AmiChat*, et *The Pacific Grove Multilingual Chat Room* [DR01], ou encore *Language Grid Playground*[21], *ChatSphere*[22] d'Apptek et *Qopuchawi*[23], le système gratuit d'Atamiri (cf. figures 2.25 et 2.26). Tous fonctionnent selon le même principe : un logiciel de tchat traditionnel est simplement couplé à un logiciel de traduction automatique qui traduit les messages. L'utilisateur spécifie, généralement à la connexion ou dans un profil, la langue qu'il écrit et la langue qu'il souhaite lire ; chaque utilisateur ne voit alors que les traductions qu'il a demandées. Ces systèmes se résument à l'assemblage d'un composant de tchat classique et d'un composant de TA tout aussi classique, sans aucune adaptation du point de vue leur ergonomie, et se révèlent donc très rudimentaires. Par exemple, aucun ne permet la post-édition des traductions.

3.4.2 Étude sur le tchat multilingue

Au début de notre travail, peu de systèmes existaient déjà, et ils ne permettaient pas de conserver des *logs* afin de les étudier. De plus, il était intéressant de voir quels problèmes techniques l'implémentation de tels systèmes basiques pouvait soulever. C'est pourquoi nous avons d'abord réalisé un outil de tchat multilingue équivalent à ceux du marché.

[21]Site officiel du logiciel Language Grid Playground : `http://www.langrid.org/playground/`
[22]Site officiel du logiciel Chatsphere : `http://www.aramedia.com/chatsphere.htm`
[23]Site officiel du logiciel Qopuchawi : `http://www.atamiri.cc/en/Atamkatiri/index.html`

Figure 2.25 – Fenêtre principale de Qopuchawi. L'interface est uniquement disponible en espagnol. Elle diffère d'un client de messagerie classique par la présence d'un menu de choix de la langue.

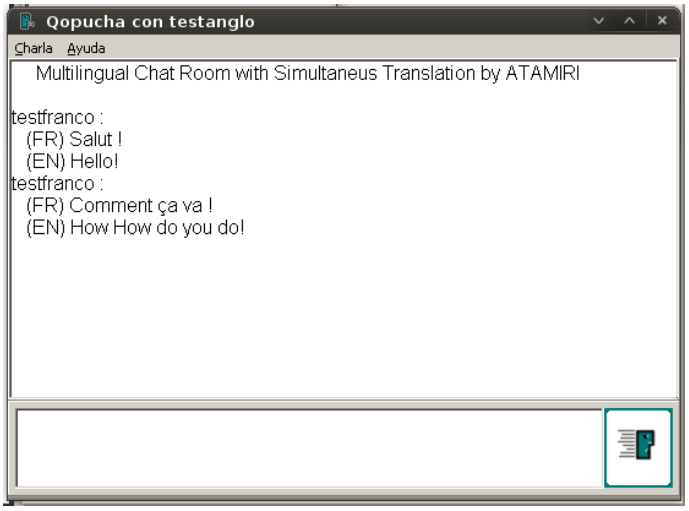


Figure 2.26 – Fenêtre de tchat bilingue d'un utilisateur francophone avec un utilisateur anglophone (l'anglais est obtenu par traduction automatique).

Différentes stratégies ont été envisagées pour assembler des composants existants de tchat et de TA.

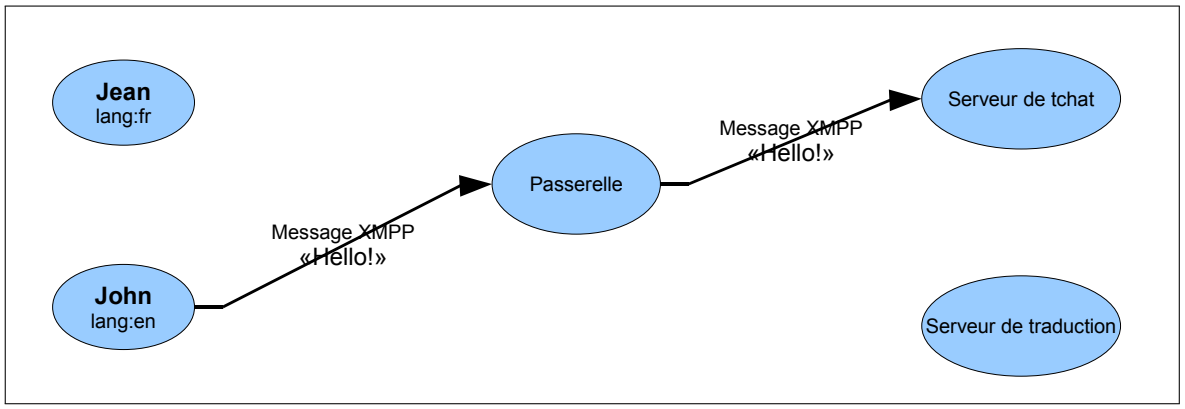


Figure 2.27 – John, un utilisateur anglophone, souhaite envoyer un message à Jean, un utilisateur francophone. Le message transite de manière transparente par la passerelle jusqu'au serveur de tchat.

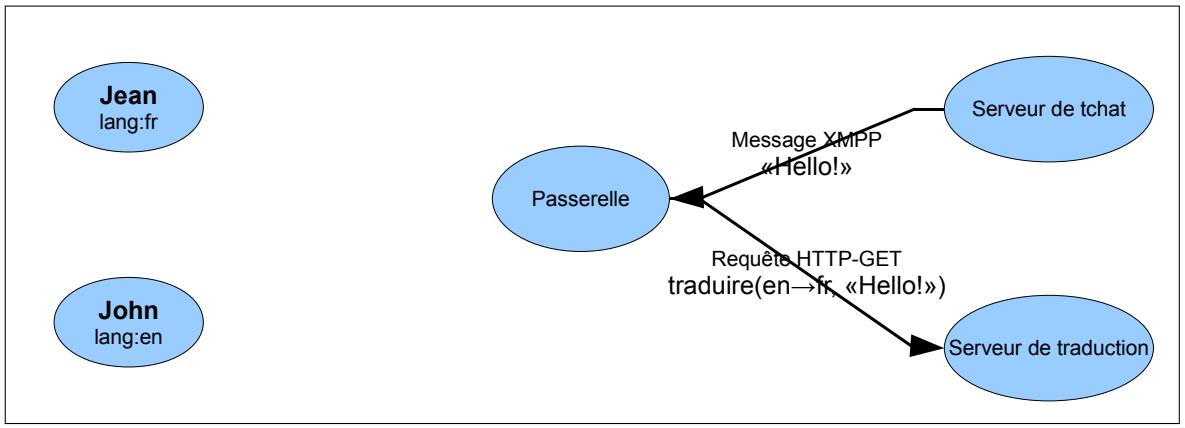


Figure 2.28 – Le serveur de tchat redirige le message de John vers Jean à travers la passerelle, mais cette dernière ne le renvoie pas immédiatement vers Jean. Le message destiné à Jean est analysé par la passerelle, qui constate que le message émane d'un utilisateur anglophone. Il est donc envoyé au serveur de traduction.

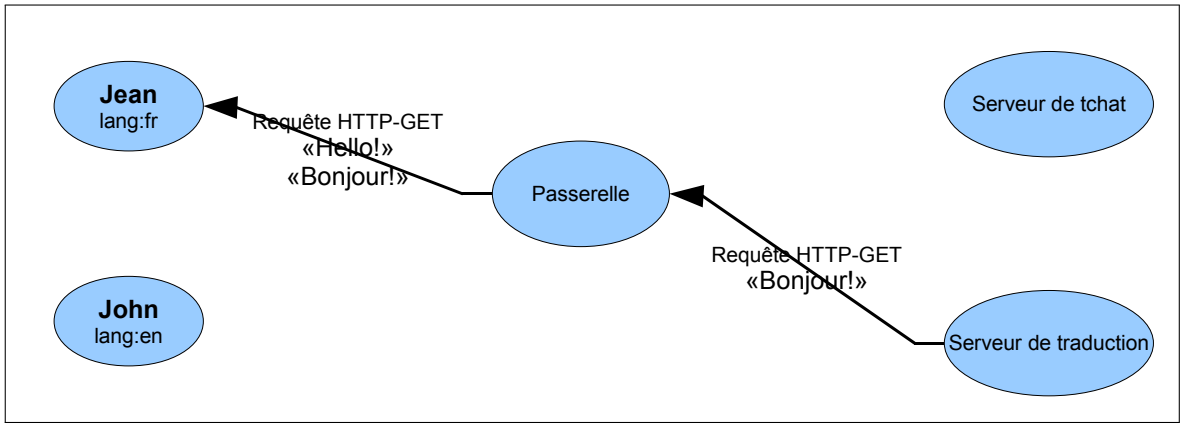


Figure 2.29 – La réponse du serveur de traduction est agrégée au message, qui peut maintenant être transféré vers Jean.

La première reposait sur le protocole IRC, et consistait à réaliser un robot. Cette implémentation présentait plusieurs avantages : d'une part, la programmation de robots se fait très simplement grâce à des langages de script ; d'autre part, le système de TA ainsi réalisé pouvait être introduit sur n'importe quel canal de tchat compatible ; il suffisait d'obtenir l'accord des usagers d'un canal pour trouver des testeurs. Mais, outre l'incertitude quant à la disponibilité et à la fiabilité des API mises à disposition par ces langages de scripts, leur statut de clients leur interdit d'aiguiller les messages. Toutes les traductions auraient donc nécessairement été adressées à tous les utilisateurs du système, y compris à ceux qui ne souhaitaient pas y avoir recours du tout. Dès lors, imaginons une discussion entre locuteurs du français, de l'anglais, de l'allemand et de l'espagnol : chaque message serait traduit dans les trois autres langues, ce qui multiplierait par quatre le nombre de messages affichés en même temps. Pour l'utilisateur, le suivi d'une telle conversation risquerait de devenir rapidement, sinon totalement impossible, du moins suffisamment pénible pour que les usagers du canal ne souhaitent pas prolonger l'expérience.

C'est finalement une implémentation de type passerelle d'accès à un serveur de tchat Jabber/XMPP qui a été retenue (cf. figures 2.27, 2.28 et 2.29). La passerelle intercepte toutes les communications entre les clients et le serveur, et modifie toutes les messages réexpédiés par ce dernier en y ajoutant leur traduction, obtenue auprès du serveur de traduction de Google. Une telle architecture simplifie l'enregistrement des conversations, et ne nécessite pas de modifier les clients (qui traitent la passerelle comme s'il s'agissait du serveur lui-même), ni le serveur (qui traite la passerelle comme si tous les clients en émanaient).

L'affichage est paramétrable, deux modes étant proposés : le mode « sous-titré » où le message original et le message traduit apparaissent tous deux, et, à titre de démonstration (étant donné le manque de fiabilité des outils de traduction automatique), le mode « doublé » où seul la traduction apparait. Le paramétrage (choix de la langue et du mode d'affichage) se fait grâce au profil utilisateur, configurable sur la plupart des clients compatibles avec le protocole Jabber/XMPP.

Les figures 2.30 et 2.31 présentent une courte conversation entre un francophone et un anglophone utilisant le client GAIM, telle qu'elle apparaît dans leurs interfaces respectives. Le serveur Jabber utilisé est eJabberd.

Conclusion

En premier lieu, on voit que tous ces outils sont confrontés au problème de l'interprétation d'énoncés ambigus, ce problème étant encore renforcé lorsque la reconnaissance vocale entre en jeu. Différentes stratégies sont utilisées. On peut utiliser une liste d'énoncés fermée (livres de phrases), en se limitant à des traductions validées, contenant éventuellement des variables lexicales. Même lorsque les énoncés sont ouverts, le système se limite souvent à un sous-domaine très limité, aussi bien du point de vue du domaine que de la finalité.

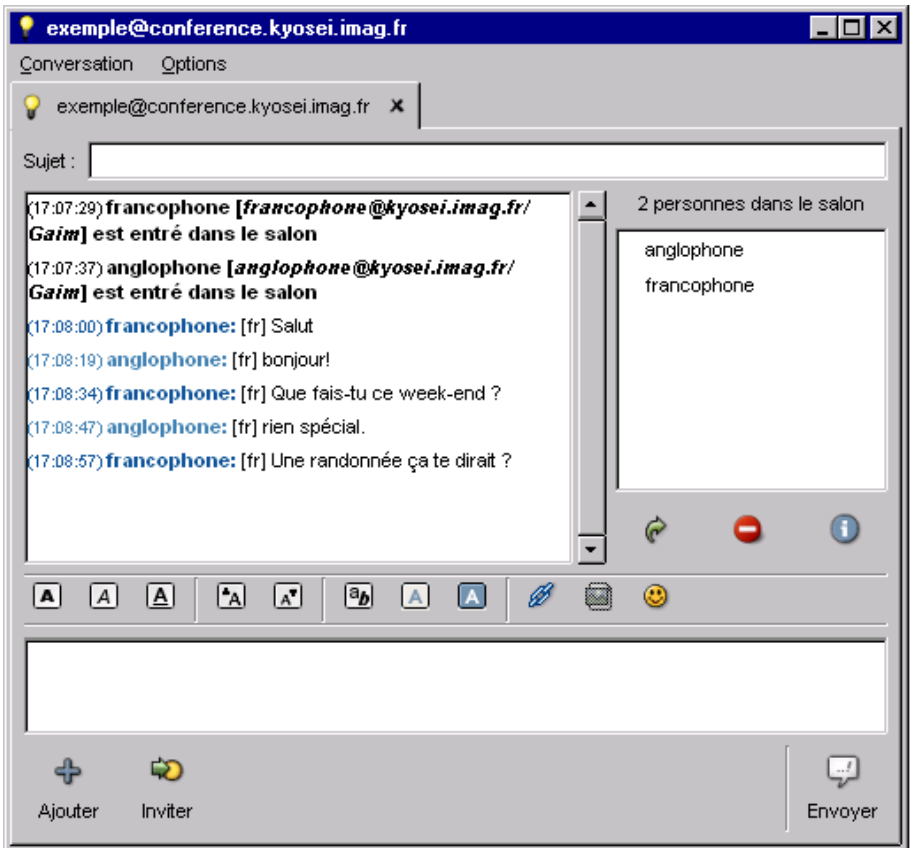


Figure 2.30 – Exemple de conversation en mode « sous-titré ». On s'y place du point de vue d'un utilisateur francophone, dialoguant avec un anglophone. Les propos de ce dernier sont traduits en français.

Parfois le domaine est totalement ouvert (tchat multilingue), mais dans ce cas la faible qualité de traduction oblige à une utilisation en « sous-titre », et encore, cela reste difficilement utilisable. En effet, pour obtenir une bonne qualité, il est nécessaire d'interagir avec le système. L'interaction la plus simple se limite à demander à un contrôle *a priori*, en demandant à l'utilisateur de valider les résultats automatiques (reconnaissance vocale *via* une transcription, ou traduction *via* une rétrotraduction).

Les systèmes plus évolués, où la fiabilité des transcriptions est importante, implémentent des fonctionnalités de désambiguïsation interactive. Les traductions sont alors de meilleure qualité, mais cette interaction a un prix, et c'est la fluidité du dialogue qui en paie la plus grande part.

Il n'est jamais prévu d'intervention de l'interlocuteur dans la construction des énoncés, un procédé pourtant courant en communication en langue seconde, comme dans toute situation de communication difficile, où l'interlocuteur tend naturellement à aider son partenaire. L'interlocuteur pourrait pourtant aider à résoudre certaines ambiguïtés, lorsque le contexte rend un choix évident. Même en se restreignant à un dialogue par tours de parole successifs, c'est à dire chaque locuteur composant ses messages seuls sans intervention extérieure, les problèmes de compréhension sont

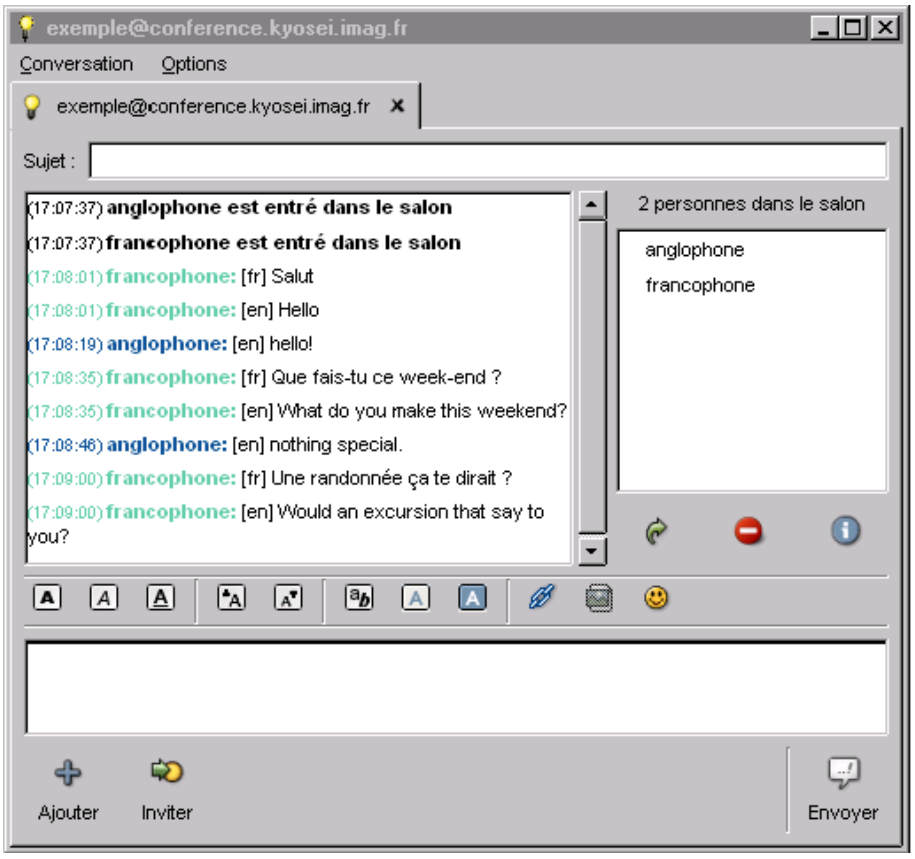


Figure 2.31 – La même conversation, en mode « doublé », du point de vue de l'interlocuteur francophone.

majoritairement relevés par l'interlocuteur, qui devrait pouvoir les signaler par le biais du système.

Par contre, les émotions sont totalement négligées par ces systèmes (à l'exception notable du projet Verbmobil qui incluait une première ébauche de reconnaissance de la colère [BHN+00]). On peut supposer que leurs utilisateurs, à l'instar des tchateurs avec leur phonétisme, leurs déformations graphiques, et leurs émoticons, vont trouver un moyen de communiquer tout de même ce type d'information.

Enfin, on voit que tous ces outils se placent d'un point de vue d'utilisateurs hétéroglottes ne maîtrisant pas de langue commune. On remarque notamment l'impossibilité pour l'utilisateur de modifier directement les traductions du système, puisqu'il est supposé totalement incompétent en la matière, alors que cela serait aisé à implémenter.

L'utilisation d'outils d'assistance est alors très utile : même si leur utilisation est coûteuse en termes de médiation, l'utilisateur s'adressant en fait à la machine qui s'adressera à son tour à l'interlocuteur, ces outils sont utilisés car la situation est bloquée. Ce genre de situation existe, en particulier en situation de crise, où l'hétéroglossie est subie par les locuteurs.

Mais comme on l'a montré au début de cette partie, les situations où une langue véhiculaire est disponible sont nombreuses. Bien sûr, des outils conçus pour une situation de blocage complet du dialogue restent utilisables dans ce contexte, en particulier si la langue véhiculaire est très mal maîtrisée par l'un des locuteurs. Mais on a observé que, même lorsqu'il conversent dans une langue qui ne leur est pas familière, les locuteurs établissent leur stratégie de signalement/réparation de problèmes de manière à minimiser les perturbations du dialogue, et à maintenir un niveau de fluidité et d'empathie élevé. Ces outils, qui imposent un coût d'utilisation à la fois fixe et élevé, sont donc fréquemment abandonnés par les locuteurs ayant à disposition une langue véhiculaire, car ils préfèrent s'en remettre entièrement à leur connaissance de la langue et à l'intelligence de leur interlocuteur.

Chapitre 3

Quels outils pour le dialogue en langue seconde ?

1 Enjeux

Les projets récents de traduction automatique de dialogue, présentés au chapitre précédent, tendent à montrer que la traduction de dialogues en temps réel, avec reconnaissance vocale, traduction automatique et enfin synthèse vocale, est aujourd'hui maîtrisable, à condition de se limiter à un sous-domaine d'application restreint. Dans ce cas, en effet, l'intelligence des locuteurs et le caractère fermé de la tâche suffisent généralement pour faire progresser le dialogue.

Mais dès que l'on cherche à étendre le domaine d'application de tels systèmes, on rencontre un taux d'erreur beaucoup trop important. En effet, beaucoup de systèmes font le choix de ne prendre en compte que l'interprétation considérée comme la plus probable, d'abord lors de la reconnaissance vocale, puis lors de l'analyse morpho-syntaxique, ce qui accroit encore un risque d'erreur déjà bien présent. Il n'est pas donc pas possible de proposer de but en blanc une telle traduction à l'utilisateur. Il faut bien voir que, même en supposant que l'on dispose d'un très bon système de reconnaissance vocale, avec un taux de réussite de 0,8 (taux aujourd'hui constaté sur des dialogues simples en domaine restreint), il s'agit d'un taux de réussite par mot. Pour un tour de parole de 5 mots, la probabilité d'une reconnaissance correcte tombe à $0{,}8^5$=0,33. Si l'on applique en plus un taux de 0,8 aux chances de succès de la traduction automatique du tour de parole, on tombe à 0,33*0,8=0,26, soit 3 chances sur 4 d'avoir un mauvais résultat en bout de chaîne.

De plus, la traduction automatique actuelle est majoritairement orientée vers la traduction de document écrits. Or, d'une part, la langue parlée est différente de la langue écrite, notamment au niveau de sa syntaxe et de sa structure [BBBD+79] ; et d'autre part, le dialogue se distingue nettement du document (ou plutôt du discours, dans une modalité orale) [Jak63] [Jak73] [Ada92], notamment par l'importance de ce que Jakobson nomme la fonction phatique du langage : dans un dialogue, de nombreuses expressions, voire des phrases entières, ne servent qu'à « tester » le

canal oral, et même plus loin, l'état d'esprit de l'interlocuteur (l'« empathie » chère à Rogers), voire son humanité.

Il est possible, en interagissant avec l'utilisateur, d'améliorer nettement ces taux de réussite [Bla94]. D'un côté, imposer de l'interaction pénalise l'ergonomie du système. Mais d'un autre côté cela se traduit par une amélioration des services rendus, tant en qualité qu'en quantité (si la reconnaissance et la traduction sont excellentes, on peut par exemple introduire aisément une synthèse vocale en sortie). Ces deux facteurs opposés, interaction et ergonomie, jouent sur la qualité globale du service, sans qu'il soit possible d'affirmer a priori quel combinaison offre le meilleur compromis.

D'autre part, les utilisateurs potentiels de systèmes de conversation multilingue maîtrisent généralement une langue commune, l'anglais par exemple, et conversent donc tout naturellement dans cette langue. Souvent, ils préfèrent finalement se fier à leur propre connaissance, même imparfaite, d'une langue pivot, plutôt qu'à un système jugé a priori, soit trop peu fiable, soit invérifiable, dans le cas où la traduction est envoyée à l'interlocuteur sans intervention de l'utilisateur (le locuteur).

Enfin, se pose le problème de la médiation : le fait de substituer un énoncé traduit à l'énoncé original entraîne une dépersonnalisation importante de la communication, une perte de la forme au détriment du fond. Ce filtrage peut s'avérer difficilement acceptable dans le cas d'énoncés oraux et d'un contexte de dialogue, dans lequel l'expression de soi, de son vécu (fonction expressive de la communication selon Jakobson, empathie de Rogers), et les jeux formels (fonction poétique) et rhétoriques qui les véhiculent souvent occupent une place importante.

Le tchat est un exemple très révélateur de ce besoin. Les utilisateurs ont rapidement cherché à faire passer des informations de ce type malgré une médiation typographique qui ne s'y prêtait guère, et ce y compris dans le cas de discussions techniques [Pie03b] [Fal04] [Fal05]. La traduction automatique, telle qu'elle existe aujourd'hui, risque de représenter un filtre encore plus strict à l'expression de ces informations ; dans le cadre du tchat, elle semble avoir été un frein important au développement des systèmes de tchat multilingue, tels que Wordlingo Tchat, ou encore Qopuchawi.

Ce besoin de minimiser la médiation est bien présent si l'on observe les interprètes d'hier (cf. figure 3.1) et d'aujourd'hui (cf. figure 3.2), qui ne se placent jamais directement entre les interlocuteurs, mais en retrait, voire derrière eux.

2 Contextes d'utilisation

2.1 Situations

Dans le cadre de cette thèse CIFRE, et en tenant compte des activités de Prosodie et de la présence de filiales à l'étranger (Espagne et États-Unis), le type de

Figure 3.1 – La Malinche traduisant le langage des *mexicas* à Cortés (source : Wikimedia, d'après Lienzo Tlaxcala, peintre du XV^{e} siècle).

Figure 3.2 – « Loge » des interprètes du parlement européen.

situation privilégié est celui de la discussion d'affaires téléphonique. Il s'agit donc d'une discussion en anglais, soit avec un seul des deux locuteurs qui n'est pas parfaitement anglophone, soit avec deux locuteurs imparfaitement anglophones, de langue maternelle différente, et employant l'anglais comme langue de communication.

Le type de discussion peut être varié (rendez-vous, commandes, etc.). Ces si-

tuations sont proches des scénarios étudiés dans le cadre du projet Verbmobil : planification de trajet, prise de rendez-vous et planification de conférence, comprenant des sous-dialogues de clarification avec l'interprète/système [JvH00]. Il s'agit d'un contexte où des problèmes de communication sont effectivement constatés, et en pratique, pour des raisons linguistiques, un passage sur support écrit (courriel) est nécessaire pour approfondir les conversations.

2.2 Langues

Ainsi, la langue par défaut sera l'anglais. Il est toutefois prévu que les intervenants puissent spécifier une autre langue, voire plusieurs langues.

En dialogue bilingue, les locuteurs peuvent proposer chacun une langue différente. Cette situation se rencontre souvent avec des apprenants, par exemple entre « correspondants » scolaires, ou dans le cadre de familles plurilingues, mais aussi dans toute situation dès lors que l'un des locuteurs maîtrise bien une langue sur la plan de la réception, mais pas en émission (on parle parfois de maîtrise « passive » de la langue).

Toutefois, il n'est pas rare de voir un dialogue en langue étrangère incorporer des éléments multilingues. Cela peut se produire involontairement, surtout lorsque les deux langues ont un vocabulaire proche. Par exemple, dans une conversation en anglais, un francophone risque d'inventer des termes anglais à partir du français (« *franglish* »), ou d'introduire des termes d'une langue proche connue (par exemple l'allemand) en pensant que c'est de l'anglais. Cela peut aussi être volontaire, pour l'expression d'une notion propre à une langue connue des deux interlocuteurs, ou par empathie, par exemple lors des salutations.

C'est pourquoi, outre la langue du dialogue, il est utile de connaître la langue maternelle des utilisateurs, mais aussi *toutes* les autres langues qu'ils maîtrisent ainsi que leur niveau.

Enfin, on aura aussi besoin de connaître la disposition du clavier de l'utilisateur.

2.3 Évaluation de la compétence linguistique

Comme nous l'avons vu, il est important de connaître le niveau de maîtrise des participants. En effet, certaines aides pourront être modulées en fonction de ce niveau. On peut utiliser un système classique, qui consiste à demander aux locuteurs de se situer eux-mêmes sur une échelle informelle, tel que celui cité en table 3.1. Toutefois, une telle auto-évaluation est trop informelle pour servir de base fiable à un système automatisé, tout au plus peut-elle permettre d'initialiser le niveau de compétence linguistique. Une mesure plus objective consiste à évaluer la couverture lexicale du locuteur [CH04]. Pour ce faire, nous comparons les fréquences des termes

Bilingue	Peut utiliser l'anglais dans toutes les situations professionnelles. Compréhension totale de conversations complexes. Prononciation, intonation et rapidité comme un autochtone.
Courant	Peut participer à des négociations et des réunions et donner des instructions dans son propre domaine professionnel. Est capable de rédiger une lettre d'affaires correcte.
Lu, écrit, parlé	Peut donner des informations concrètes ou des instructions dans le milieu professionnel. Peut comprendre une conversation téléphonique en faisant parfois répéter.
Scolaire	Peut donner des instructions simples sans discussions ni commentaires. Ne peut pas être impliqué de manière autonome dans des situations professionnelles en anglais. Peut difficilement suivre une conversation téléphonique. Expression écrite plus claire que l'expression orale.
Notions	Ne peut pas être impliqué de manière autonome dans des situations professionnelles en anglais. Comprend le contenu de questions simples.

Table 3.1 – Échelle des niveaux d'anglais pour les CV, selon le site www.go.tm.fr

utilisés par le locuteur (vecteur L) à un modèle thématique de référence (vecteur R), à l'aide d'une similarité cosinus (cf. équation 3.1).

$$niveau(L|R) = ArcCos\frac{L \cdot R}{||L|| \cdot ||R||} \tag{3.1}$$

La mesure ainsi calculée est exprimée en degrés d'arc, et appartient donc à l'intervalle [0..180]. Une valeur élevée signifie une couverture lexicale différente de celle d'un locuteur natif. On remarquera toutefois que ce type de mesure semble insuffisant pour les locuteurs non natifs mais néanmoins expérimentés [Tho08].

Ainsi, au delà du problème d'initialisation de ce paramètre, l'évaluation de la compétence pourra ensuite être automatiquement ajustée, voire déduite si l'utilisateur a omis de la spécifier, en fonction des productions des utilisateurs.

3 Les aides

Afin de répondre à ces problèmes de communication, le système d'aide envisagé prend en charge les modalités orale et écrite, dans la perspective d'une utilisation en cascade : la modalité vocale est utilisée par défaut, et les locuteurs passent à l'écrit en cas de difficulté. En outre, un système de collaboration graphique est prévu, permettant aux participants d'associer à la conversation des images et des pages Web, et de les annoter collaborativement (cf. figure 3.3).

On l'a vu, il est important de limiter autant que possible la médiatisation de la communication. D'un autre côté, l'interaction personne-système, et donc la médiatisation, est indispensable, à divers degrés, dès lors que l'on souhaite utiliser des

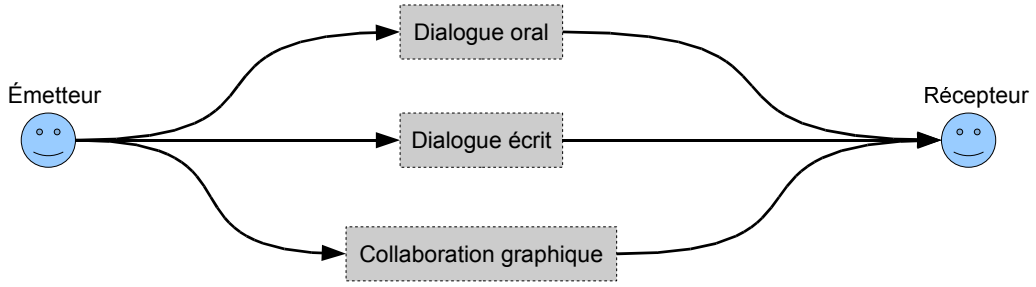


Figure 3.3 – Modalités proposées.

composants d'assistance linguistique. Il y a donc un compromis à trouver, en fonction des difficultés rencontrées par les locuteurs. Or, bien sûr, au fil de la conversation, ces dernières ne sont ni continuelles, ni toujours les mêmes ; les aides doivent donc être à disposition, mais jamais contraignantes ou bloquantes. Pour ce faire, nous proposons d'étendre au sein de chaque modalité l'organisation en cascade, avec des aides passives et, sur l'initiative de l'utilisateur, des aides actives. Ainsi, le système ne ralentit pas la conversation, qui reste synchrone autant que possible : il est possible d'ignorer totalement les aides tant que tout se passe bien.

En cas de problème, l'utilisateur peut donc s'appuyer sur des aides actives et passives. Les aides actives vont lui servir à améliorer sa production, et les aides passives à l'évaluer. En réception, à l'écrit, une aide active permet à l'interlocuteur de signaler des passages difficiles et d'entamer un sous-dialogue de clarification. Quant aux informations fournies par les aides passives, elles vont faciliter sa compréhension.

Mais que faire lorsque, grâce à une aide passive, l'utilisateur repère un problème potentiel ? On pourrait bien sûr le laisser corriger le problème directement, mais il est souvent plus intéressant d'associer à chaque outil de mesure passif un outil de correction actif.

C'est au total plus simple pour l'utilisateur, et partant de là, plus rapide, et donc bénéfique pour la synchronicité du dialogue. D'autre part, cela permet de conserver un historique détaillé des modifications et de leurs motivations. En outre, les informations données par les aides passives dénotent la « compréhension » du dialogue par le système. Toutefois, ce dernier peut se tromper, et fournir des informations erronées. L'utilisateur peut alors corriger ces informations, et cette correction est alors prise en compte par le système pour améliorer cette « compréhension ».

Le dialogue s'accompagne alors, outre des informations calculées par les aides passives, des réactions circonstanciées des utilisateurs eux-mêmes et de leurs modifications. En intégrant un suivi des modifications, la mémoire de dialogue ainsi obtenue emprunte aux mémoires de traductions et aux documents auto-explicatifs [BB04b] [BB04a], en ajoutant des éléments propres à l'aide au dialogue en langue

étrangère.

Dans le cadre de cette thèse, nous ne cherchons pas à mettre en place les meilleures aides possibles. Nous privilégions la question de l'harmonisation des approches et la simplicité d'implémentation, dans la perspective d'une expérimentation globale. De plus, quelle que soit la qualité des aides apportées, ces aides peuvent se révéler inadaptées, ou en tout cas moins efficaces que l'intelligence des locuteurs. Nous avons souhaité intégrer ce problème. C'est pourquoi, dans le cas des modalités graphiques (écrit, collaboration graphique, *via* les aides pour la parole), les hypothèses du système sont visibles et corrigeables par les locuteurs. Par ailleurs, les énoncés sont post-éditables par leur auteur, et d'autre part, annotables directement (signalement de passages difficiles). Lorsqu'un utilisateur note une difficulté *via* le système, il peut alors entamer un sous-dialogue de clarification avec l'interlocuteur en cause.

Ce modèle d'interaction global est schématisé sur la figure 3.4.

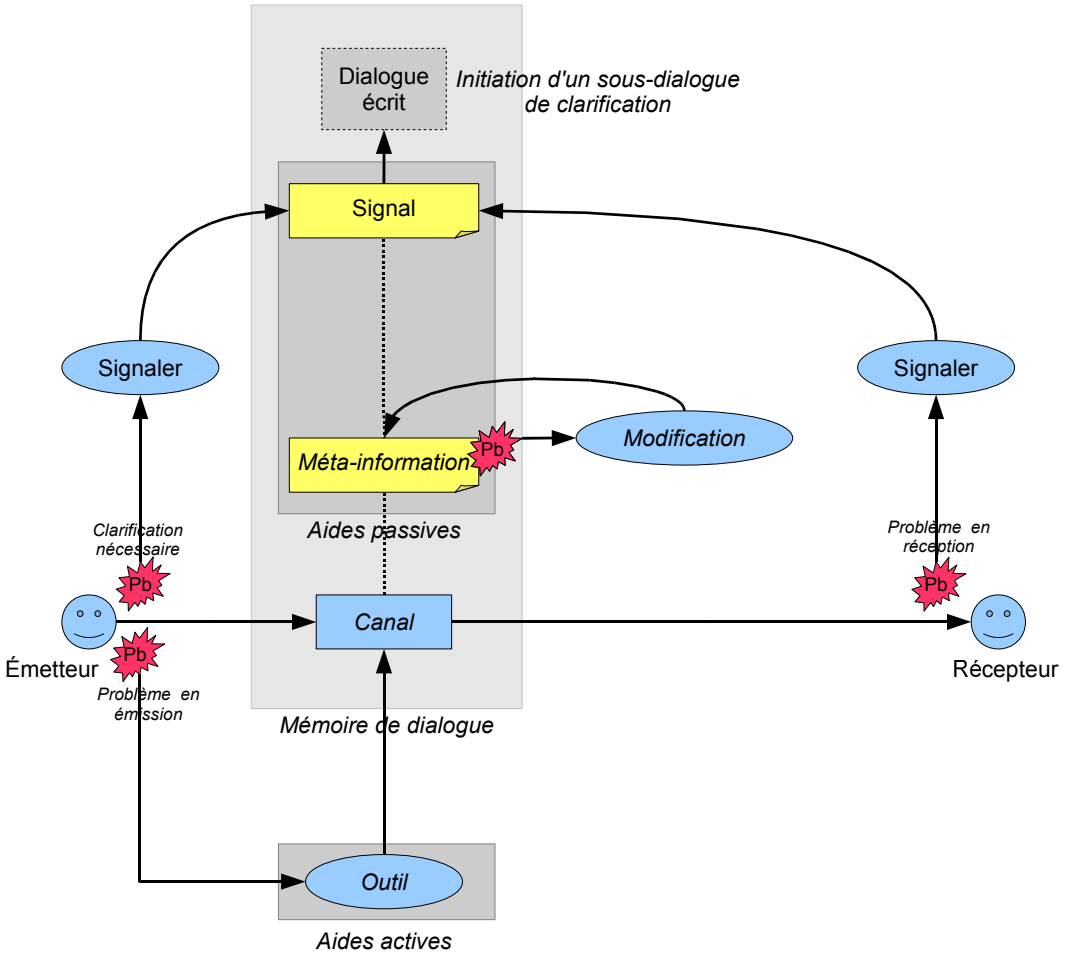


Figure 3.4 – Modèle d'interaction de Koinè.

Étant donnés les problèmes techniques engendrés par la reconnaissance vocale, nous n'avons pas souhaité trop développer une modalité qui, dans le cadre de la thèse, restera au stade exploratoire. Nous donnerons juste un aperçu de ce que pourraient être des aides pour cette modalité, sans entrer dans les détails, en particulier en ce qui concerne les algorithmes utilisés : à ce niveau, nous en restons encore malheureusement à l'approche dominante, qui consiste à transformer l'oral en texte avant de le traiter. On notera toutefois que le texte en question comporte une dimension informelle assez inédite dans un tel système, et qu'un système traitant la parole spontanée de façon native devra nécessairement intégrer cette dimension. L'écrit, qui présente nettement moins de problèmes techniques, sera présenté de façon beaucoup plus détaillée.

Certaines des mesures présentées pour cette modalité conviendraient tout à fait à de la parole transcrite, mais dans la mesure où cela revient à transposer des outils de l'écrit à la parole, nous avons préféré les présenter dans leur modalité d'origine. Enfin, nous présenterons la modalité graphique, qui présente surtout un intérêt par son intégration au système de dialogue.

4 Mode vocal

L'état de l'art en reconnaissance vocale, pour des conversations téléphoniques spontanées entre locuteurs non natifs, limite fortement les applications possibles. Cela ne nous a pas empêché de réfléchir, compte tenu des ces limitations techniques, mais aussi des avancées récentes, aux aides que l'on pouvait apporter dans le cadre de la modalité vocale. Néanmoins, bien qu'hypothétiquement possibles, leur mise en œuvre n'a rien de trivial et soulève encore bien des problèmes techniques sortant du cadre de cette thèse. C'est pourquoi les aides proposées au niveau vocal doivent bien être comprises comme étant à visée purement exploratoire. Le chapitre 5, qui traite de l'implémentation, reviendra plus en détail sur les avancées et les problèmes propres à la reconnaissance vocale.

Nous présenterons dans un premier temps les principaux problèmes liés à l'utilisation de reconnaissance vocale dans notre contexte d'utilisation, puis quelques outils envisageables en dépit de l'imperfection de la reconnaissance vocale.

4.1 Problèmes et pistes de recherche

Le problème global soulevé par ce projet concerne l'adaptation des outils de reconnaissance vocale et de traduction automatique au dialogue téléphonique spontané. C'est un problème très vaste, où l'on peut distinguer plusieurs sous-problèmes plus spécifiques. Ces problèmes se retrouveront à l'écrit, sous une forme légèrement différente.

4.1.1 Incertitudes et discontinuités

Quelle que soit l'acuité de la reconnaissance vocale, certains passages poseront toujours problème : si parfois les humains eux-mêmes ont besoin qu'on leur répète

une phrase, il ne faut pas s'attendre à ce qu'une machine comprenne toujours chaque production orale sans aide. Étant donné notre souci de ne pas perturber la spontanéité du dialogue, il n'est pas possible d'interrompre sans arrêt le locuteur pour lui demander de répéter (cf. par exemple le projet C-STAR). Nous devons donc travailler avec décodage de base plus ou moins fiable, et donc au niveau des données de base (la parole reconnue), nous concentrer sur des îlots de très relative certitude, au lieu de traiter un flux constant.

Avec ces énoncés discontinus, un problème se pose : il faut sélectionner les îlots suffisamment complets pour constituer des blocs autonomes, en tentant lorsque c'est possible d'établir un lien entre les îlots, à l'aide d'une mémoire de traduction à court terme.

4.1.2 Adapter la reconnaissance vocale et la traduction automatique

D'autre part, la désambiguïsation interactive, lorsqu'elle est présente, doit se limiter aux questions pertinentes pour la langue cible. En effet, celle-ci alourdit notablement la communication. Il ne s'agit pas, à ce niveau, de générer des documents auto-explicatifs [BB04a] exhaustifs, mais de lever les ambiguïtés vraiment problématiques. Certaines ambiguïtés, sensibles en langue cible mais n'ayant pas d'impact trop important sur la compréhension, pourraient ainsi ne pas être levées, en particulier quand une alternative apparaît nettement plus probable que les autres.

Par conséquent, il faut gérer une information ambiguë, lorsqu'en l'absence de désambiguïsation interactive (en réception ou bien si le service a été désactivé), une partie du signal de parole n'a pu être reconnue.

De plus, les énoncés peuvent aussi être eux-mêmes multilingues. Outre le problème de modélisation que cela implique, se pose le problème de l'adaptation à un système de traduction automatique qui prend en entrée une information simple, « plate ». Une possibilité est d'adapter l'information sous forme de variantes avant de les soumettre au module de traduction automatique, et de recomposer ensuite une version factorisée, à nouveau ambiguë.

4.1.3 Adaptation des outils au contexte téléphonique

D'un point de vue ergonomique, le nécessité d'un bouton « push to talk » pose elle aussi problème. Ce n'est pas un type d'interaction naturel pour un simple utilisateur de téléphone, et la contrainte supplémentaire (réelle dans le cadre d'un dialogue) représentée par ce bouton risque de n'en être que plus mal perçue. Sans compter qu'une fois le bouton relâché, il faut encore attendre que le traitement automatique prenne fin, puis que le message soit transmis à l'interlocuteur. A ce niveau, un peu plus de fluidité serait bienvenue. Or, si les systèmes de reconnaissance vocale sont relativement bien adaptés au traitement de la parole continue, cela pose plus de problèmes au niveau de la traduction automatique. La conception d'un système de traduction traitant une information continue, non-bornée, semble difficile ; mais on

peut envisager de stocker l'information dans un tampon, et de soumettre ainsi des segments relativement autonomes au système. Cela pose, d'une part, un problème de segmentation, pour lequel l'information prosodique pourrait être utilisée avec profit (silences, frontières prosodiques marquées), et d'autre part, de mémoire de traduction à court terme, afin de ne pas perdre le contexte propositionnel.

4.1.4 Gestion des accents et des langues

Il faut tenir compte du fait qu'au moins un des locuteurs ne s'exprime pas dans sa langue maternelle, ce qui peut donner des accents très variables (penser par exemple à de l'anglais avec l'accent du midi !), et perturber sérieusement le système de reconnaissance vocale. À défaut d'un système adapté, coûteux à développer, une piste concerne le couplage d'un modèle de parole de langue 1 avec un modèle de langage de langue 2 (adapté aux phonèmes de la langue 1), comme le montre une thèse sur le sujet [TB06], ou plusieurs travaux concernant les langues peu dotées [Ber04] [BLC+05] [Lê06].

En réception comme en émission, il est possible que certains énoncés soient multilingues, on a alors un problème d'identification des langues, qui vient s'ajouter aux opérations de segmentation. En utilisant, comme précédemment, et en fonctions des langues configurées par l'utilisateur, une batterie de moteurs de reconnaissance, mais aussi de traduction, on devrait pouvoir faire émerger la langue qui « fonctionne » pour tel ou tel segment de l'énoncé.

4.2 Mise en œuvre

On cherche donc à mettre en œuvre un jeu progressif de services, pouvant être activés ou non, en fonction de la configuration effectuée par l'utilisateur.

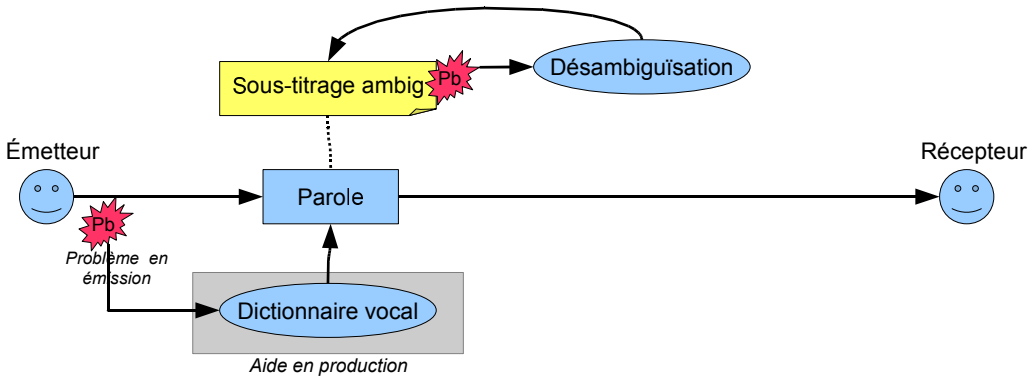


Figure 3.5 – Organisation des aides du mode oral.

4.2.1 Dictionnaire vocal

Dans sa version la plus simple, le système ne propose aucune traduction ; le dialogue se fait sans traitement, ni contrainte, ni délai, dans une langue commune aux deux interlocuteurs.

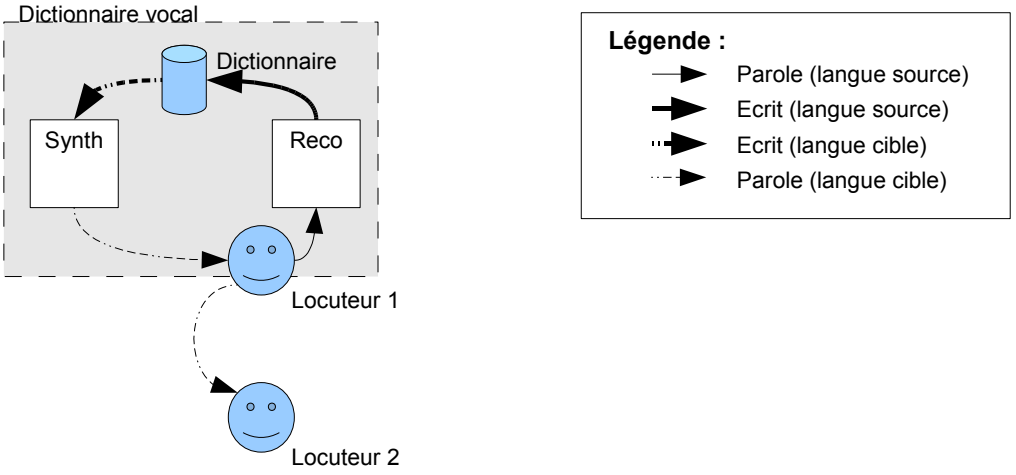


Figure 3.6 – Dictionnaire vocal

Lorsqu'il ne trouve pas un mot (problème fréquent lorsqu'on maîtrise un minimum la syntaxe d'une langue, mais que le vocabulaire reste limité), l'utilisateur peut appuyer sur un bouton, et prononcer un mot dans sa langue maternelle pendant que ce bouton est enfoncé. Ce mot est détourné par le système vers un système de reconnaissance vocale, avant d'être recherché dans un dictionnaire bilingue et synthétisé au retour.

On peut imaginer une mini-désambiguïsation interactive avant la synthèse, lorsque la reconnaissance vocale hésite entre plusieurs mots, dans le cas d'homophones, ou encore lorsque le mot donné correspond à plusieurs mots possibles en langue cible : une liste d'entrées candidates apparaît, il suffit de cliquer sur l'entrée voulue.

4.2.2 Version plus élaborée

Dans une version plus élaborée, qui nécessite d'introduire un léger différé dans la conversation, l'utilisateur parle normalement dans la langue de dialogue, mais il peut à tout moment prononcer un mot dans sa langue maternelle, et en le signalant par l'appui sur un bouton à la place d'un mot en langue cible. Ce mot sera traduit en langue cible, éventuellement désambiguïsé interactivement (en cas de forte ambiguïté), et synthétisé, avant d'être intégré dans l'énoncé en langue cible de l'utilisateur.

Bien sûr, cela implique que le mot prononcé en langue source soit clairement délimité par des silences. En langue cible, il faudra se contenter d'une prosodie « standard », et d'une forme quasi-lemmatisée, puisqu'aucune analyse morpho-syntaxique de la langue cible n'est effectuée avant l'insertion.

4.2.3 Sous-titrage en réception

Un autre service peu coûteux en termes de médiation concerne le sous-titrage du signal en réception, en version originale. Le signal vocal est intercepté mais non modifié. Un sous-titre est affiché, plus ou moins étendu selon les capacités de la reconnaissance vocale. Bien sûr, ce sous-titrage ne peut pas apparaître instantanément, et ne pourra être affiché que sous la forme de *logs*, en léger différé.

Les ambiguïtés de reconnaissance vocale sont affichées sous forme graphique, avec une mise en évidence de chaque hypothèse proportionnelle à sa probabilité. On se déplace alors dans une modalité graphique, dans laquelle les utilisateurs peuvent intervenir directement sur le graphe, soit pour confirmer/modifier/supprimer certaines hypothèses, soit pour signaler un problème, pouvant éventuellement mener à un sous-dialogue de clarification. On se retrouve en fait dans une situation de modalité écrite, avec un écrit présenté sous forme de graphe d'ambiguïtés.

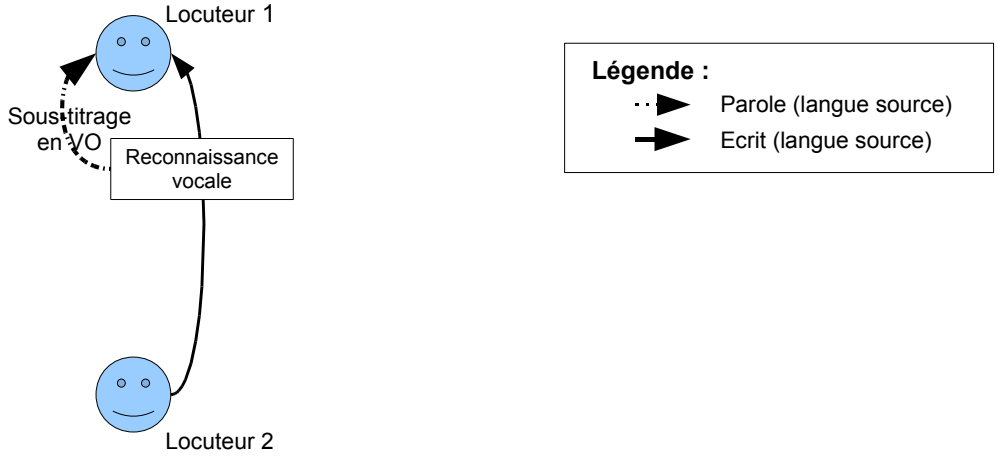


Figure 3.7 – « Sous-titrage » partiel en langue source.

5 Mode écrit

Le mode écrit pose clairement moins de problèmes de décodage que le mode vocal, mais n'en est néanmoins pas totalement exempt. S'agissant de tchat, c'est à dire d'un écrit informel, dans lequel, comme à l'oral, le temps de la planification linguistique tend à se confondre avec celui de la production, il faut s'attendre à voir

des graphies non-standard. Sans que cela perturbe nécessairement les locuteurs : leur expertise linguistique peut leur permettre de comprendre sans difficulté ces graphies, qui peuvent en outre être volontaires, et motivées par des soucis d'expressivité et d'empathie nécessaires à la conduite du dialogue dans de bonnes conditions.

Dans cette perspective, comme à l'oral, il n'est donc pas acceptable de contraindre les énoncés, en obligeant les utilisateurs à écrire « correctement ». À ce niveau, la première aide que l'on peut leur apporter est la possibilité de corriger, d'annoter et de discuter *a posteriori* leurs messages, c'est à dire simplement de donner un cadre formel à ce qu'ils font déjà malgré des outils qui ne les y aident pas.

Dans ces conditions, il est difficile d'envisager des outils « globaux » sensibles à la qualité des données, comme des outils qui se baseraient sur la structure du dialogue [Cae02] ou sur une traduction automatique, car cela obligerait les utilisateurs à respecter un cadre formel tout au long du dialogue. Par contre, des aides très locales, par exemple lexicales, peuvent fonctionner, car elles ne contraignent qu'une petite portion du dialogue, et précisément celle qui pose problème aux utilisateurs. Dans les cas où une analyse globale s'avère néanmoins nécessaire, nous privilégions des modèles statistiques tolérants à l'erreur, plutôt que des modèles heuristiques.

5.1 Vue d'ensemble

5.1.1 Fonctionnalités

La figure 3.8 donne un aperçu des aides mises en œuvre pour l'écrit.

Les aides passives peuvent être utilisées en réception comme en production. Quand un problème est repéré grâce à une de ces aides, on essaye d'offrir en même temps un outil de correction adapté. Si les utilisateurs ne sont pas d'accord avec les mesures automatiques, ils peuvent les corriger. Cette correction est visible par l'interlocuteur, qui peut réagir, et éventuellement lancer un sous-dialogue de clarification.

Signal : permet aux utilisateurs de « marquer » le texte à l'intention de leurs interlocuteurs.

Traduction automatique : permet de consulter une traduction automatique, et éventuellement de la corriger.

Sous-titrage lexical : permet de voir le texte en mode « sous-titré » : pour chaque mot (simple ou composé), sa traduction la plus probable est donnée. Le choix de la traduction la plus probable peut être corrigé, et la décision de sous-titrer ou non un mot se fonde sur la fréquence du mot rapportée au niveau du lecteur.

Intelligibilité : donne une mesure de l'intelligibilité d'un tour de parole.

Prototypicalité : marque les formules « typiques » en fonction de la langue et du thème. Lorsqu'une séquence de mots est jugée « typique », un mini-concordancier permet d'en contrôler l'usage.

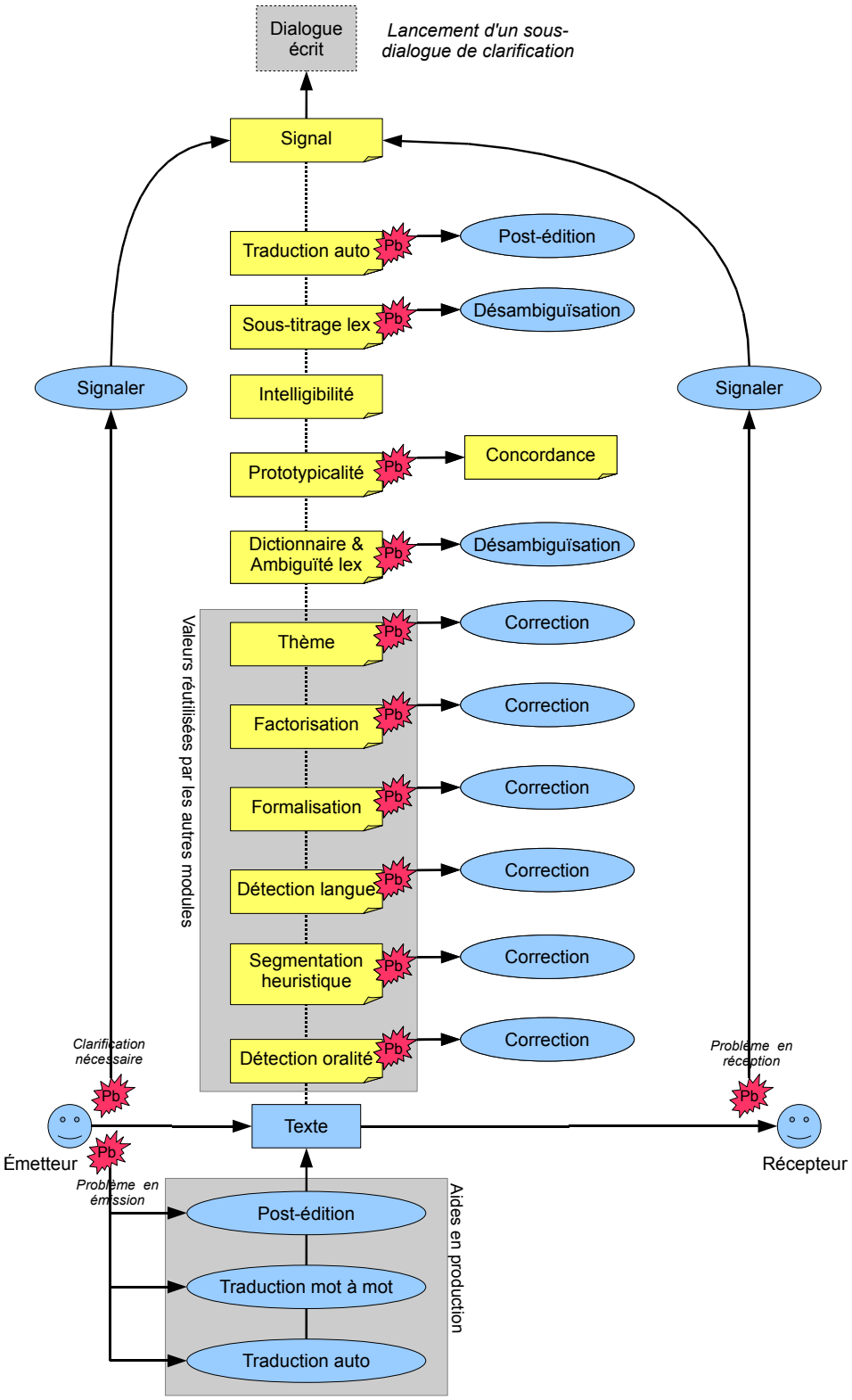


Figure 3.8 – Organisation des aides du mode écrit.

Dictionnaire & ambiguïté lexicale : indique les mots (simples ou composés) qui

peuvent être ambigus, en fonction du thème, du contexte, et surtout de la langue de l'interlocuteur. Lorsqu'une ambiguïté n'existe pas dans une langue, un contexte d'utilisation désambiguïsant est donné.

Formalisation : indique les abréviations développées automatiquement, et indique les mots inconnus ainsi que des suggestions de « corrections » sur un mode « correction orthographique ».

Pour chaque aide passive, il est possible de lancer un sous-dialogue de clarification lié à l'aide en question.

Certaines aides passives, basées sur des mesures statistiques sont activées en fonction d'un seuil, initialisé empiriquement et ajustable par l'utilisateur.

Les aides actives sont spécifiques à la production d'énoncés, et ne peuvent être utilisées que par l'auteur du « tour de parole » :

Post-édition : permet à l'auteur de modifier son texte *a posteriori*. Un historique est conservé.

Traduction mot à mot : permet à l'auteur d'entrer un mot dans sa langue maternelle, qui sera traduit automatiquement après sa saisie.

Traduction automatique : permet à l'auteur de traduire automatiquement son énoncé, puis éventuellement de post-éditer le résultat (sans historique).

5.1.2 Organisation des fonctionnalités

Comme nous l'avons vu, notre système doit pouvoir tolérer de l'écrit non standard, afin que les utilisateurs puissent continuer à échanger des éléments paraverbaux. Pour tolérer ce type d'écrit, une alternative s'offre à nous : soit nous développons des outils et des ressources spécifiques pour le tchat en langue seconde, soit nous réutilisons les outils existants, conçus autour de l'écrit standard, en essayant de «standardiser» l'écrit.

Dans la mesure où de très nombreux outils et ressources existent déjà pour l'écrit standard, c'est cette seconde solution que nous avons retenue. Nous distinguons trois catégories de fragments textuels :

- Les fragments standards, qui correspondent à de l'écrit standard et ne posent pas de problème pour le système.
- Les fragments oralisés, qui correspondent à de l'écrit standard transformé (« fautes » volontaires, comme des abréviations, des allongements, etc.) ou erroné (fautes involontaires : fautes de frappe ou d'orthographe). Ces fragments ont un équivalent standard, qu'il faudra trouver.
- Les fragments oraux, qui n'ont pas d'équivalent standard, comme les émoticons. Ces fragments devront être ignorés.

Le système pouvant se tromper dans sa classification, ainsi qu'en ce qui concerne la standardisation d'un fragment oralisé, cette tâche doit se faire sous la supervision de l'utilisateur. Toutefois, il n'est pas question d'obliger l'utilisateur à standardiser ses énoncés, puisqu'on a vu que cette dimension non standard était utile ; ce processus

de classification et de standardisation se fera donc sous formes d'annotations sur le texte original non standard, qui sera ainsi préservé.

En outre, demander à l'utilisateur d'annoter chacun de ses messages serait assez pénible pour les utilisateurs, qui ont déjà des stratégies de dialogue en langue seconde qui peuvent parfois être supérieures aux aides que nous souhaitons apporter. C'est pourquoi nous présentons ces aides sous forme contextuelle : il est nécessaire d'annoter le texte non standard avant de les utiliser, mais seulement la portions de texte sur lequel l'aide devra porter. Nous qualifions cette approche de « médiation faible », en opposition à la « médiation forte » généralement utilisée dans les applications d'aide au dialogue (voir partie 2), dans lesquels les messages doivent obligatoirement être intégralement normalisés avant le traitement.

Concrètement, l'interface s'organise en deux colonnes : la discussion à gauche, comme dans n'importe quel tchat (voir figure 3.9), et les aides à droite.

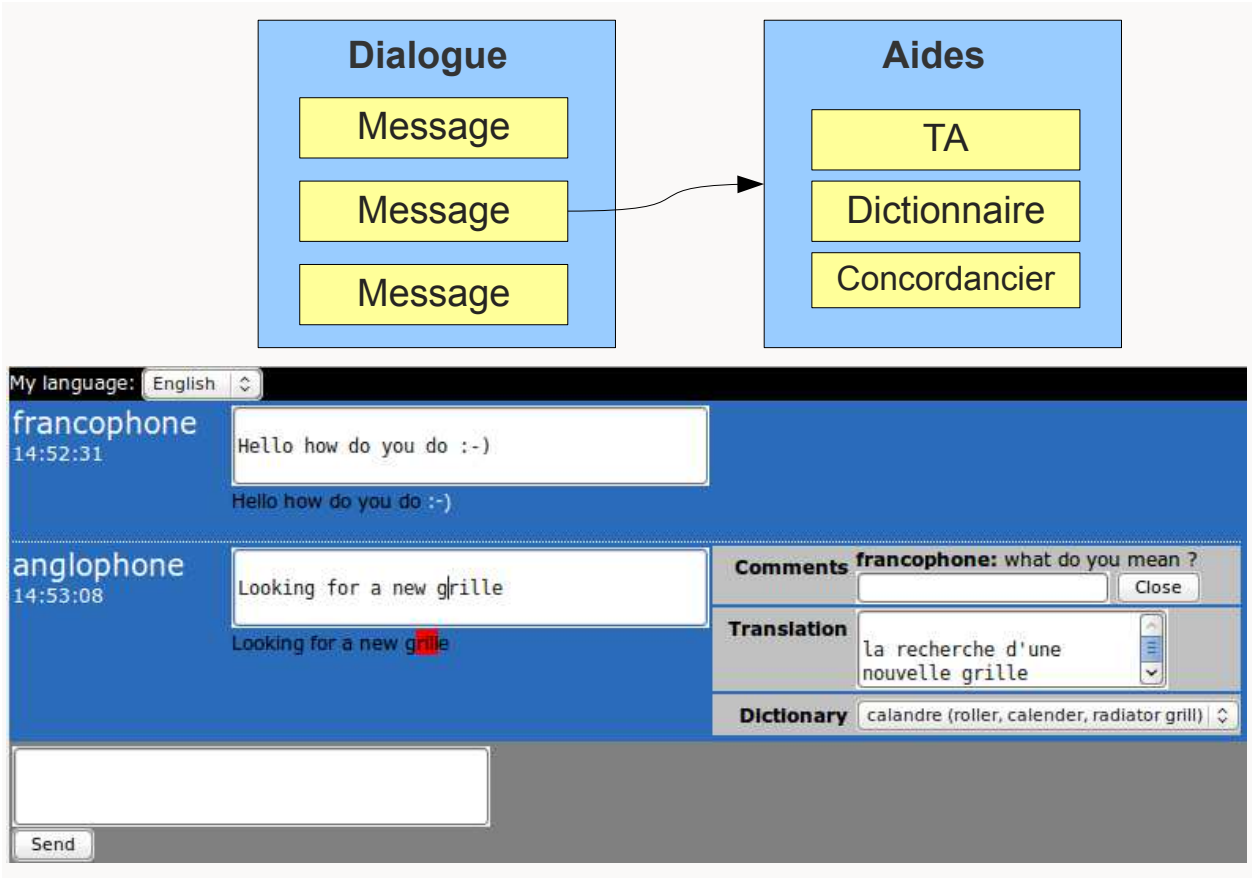


Figure 3.9 – Organisation de l'interface en médiation faible, schéma (en haut) et capture d'écran (en bas).

D'autre part, les utilisateurs ont déjà des stratégies pour ce type de dialogue. Celles-ci comportent généralement une part métalinguistique, allant parfois jusqu'au

sous-dialogue de désambiguïsation. Ce type de sous-dialogue est pris en charge par les aides (cf. figure 3.10).

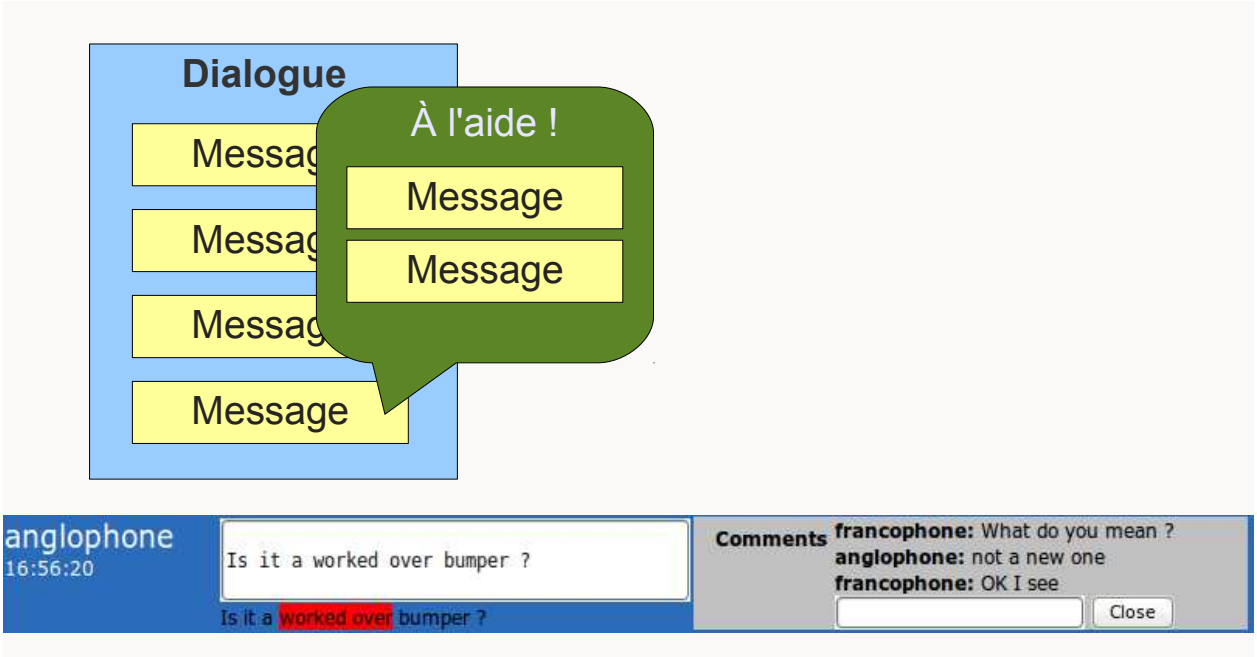


Figure 3.10 – Sous dialogue métalinguistique, schéma (en haut) et capture d'écran (en bas).

Mais le processus d'annotation lui-même est métalinguistique, et les annotations peuvent parfois servir de support au sous-dialogue (voir figure 3.11).

5.1.3 Ressources linguistiques

Dans la mesure où nous souhaitons pouvoir adapter notre outil à de nombreuses langues, nous nous basons sur des ressources homogènes et à large couverture, et donc correspondant à un plus petit dénominateur commun de caractéristiques. Ainsi, pour les dictionnaires bilingues, nous nous contentons d'une structure qui associe à un mot en langue source ses traductions possibles en langue cible. De nombreux dictionnaires libres ne donnent pas plus d'informations [1].

En ce qui concerne le modèle de langage, nous nous basons sur Wikipédia, l'une des plus importantes ressources multilingues disponibles, et en particulier sur les pages de discussion. Celles-ci sont certes assez différentes d'une discussion de tchat, mais hormis pour le français et l'italien, on ne dispose pas de corpus de langue tchatée suffisamment étendus pour construire un modèle de langage fiable. À partir des *dumps* de la base de donnée librement accessibles, nous construisons pour chaque langue des modèles de langage thématiques, trigramme, syntagmatique et paradigmatique (ces deux dernières notions seront expliquées par la suite). La construction de ces ressources sera détaillée dans la dernière partie de la thèse.

[1] *XML Dictionary Exchange Format*, `http://xdxf.sourceforge.net/`

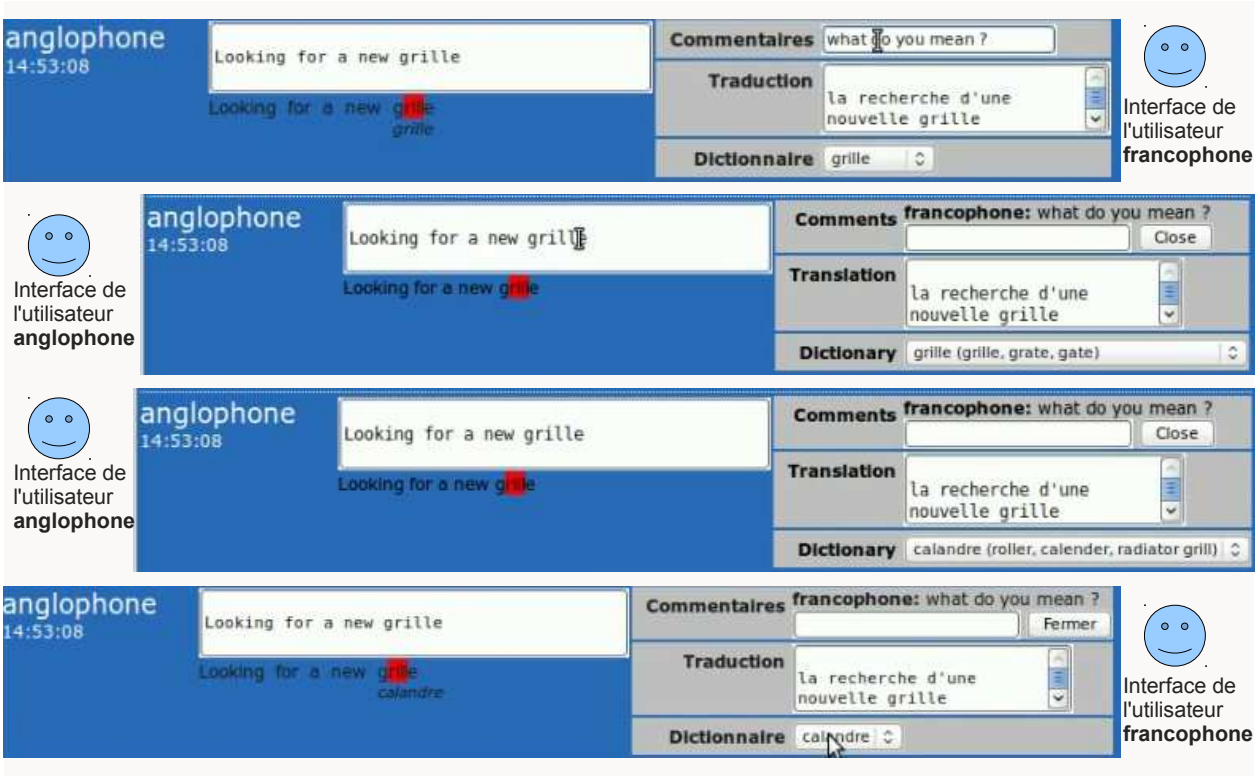


Figure 3.11 – Collaboration métalinguistique en 4 étapes sur un même message rédigé par le locuteur anglophone, dans le cadre d'une discussion sur le thème de l'automobile. (1) Le locuteur francophone s'interroge sur le sens du terme anglais *grille* ; il n'est pas satisfait par la traduction par défaut donnée par le dictionnaire contextuel (« grille ») ; il surligne donc le mot et demande à son interlocuteur anglophone de désambiguïser ce terme. (2) Cette demande est reçue par l'interlocuteur anglophone, et (3) ce dernier utilise le dictionnaire pour désambiguïser le terme (*grille* dans le sens de *roller*, *calender*, *radiator grill* : « calandre » en français). (4) L'utilisateur francophone peut maintenant voir une traduction désambiguïsée dans son dictionnaire contextuel.

5.1.4 Architecture

Le projet Verbmobil impliquait de nombreux modules de traitement, conçus par des équipes différentes et devant traiter en synergie un flux de données [KNK00]. Une première maquette s'appuyait sur une architecture multi-agents pour faire coopérer ces différents modules, les agents communiquant directement par messages, mais ne se révéla pas satisfaisante (cf. chapitre 2). La seconde impliquait des modules incapables de communiquer directement mais partageant des données, selon une architecture « tableau noir multiple ».

L'organisation des traitements dans *Koinè* s'inscrit dans cette optique, même si le nombre plus faible de modules et leur plus grande homogénéité autorisent quelques simplifications.

Structure de données Cette structure est constituée de différents niveaux, de granularité croissante. À chaque niveau, l'unité de référence se voit attacher diverses

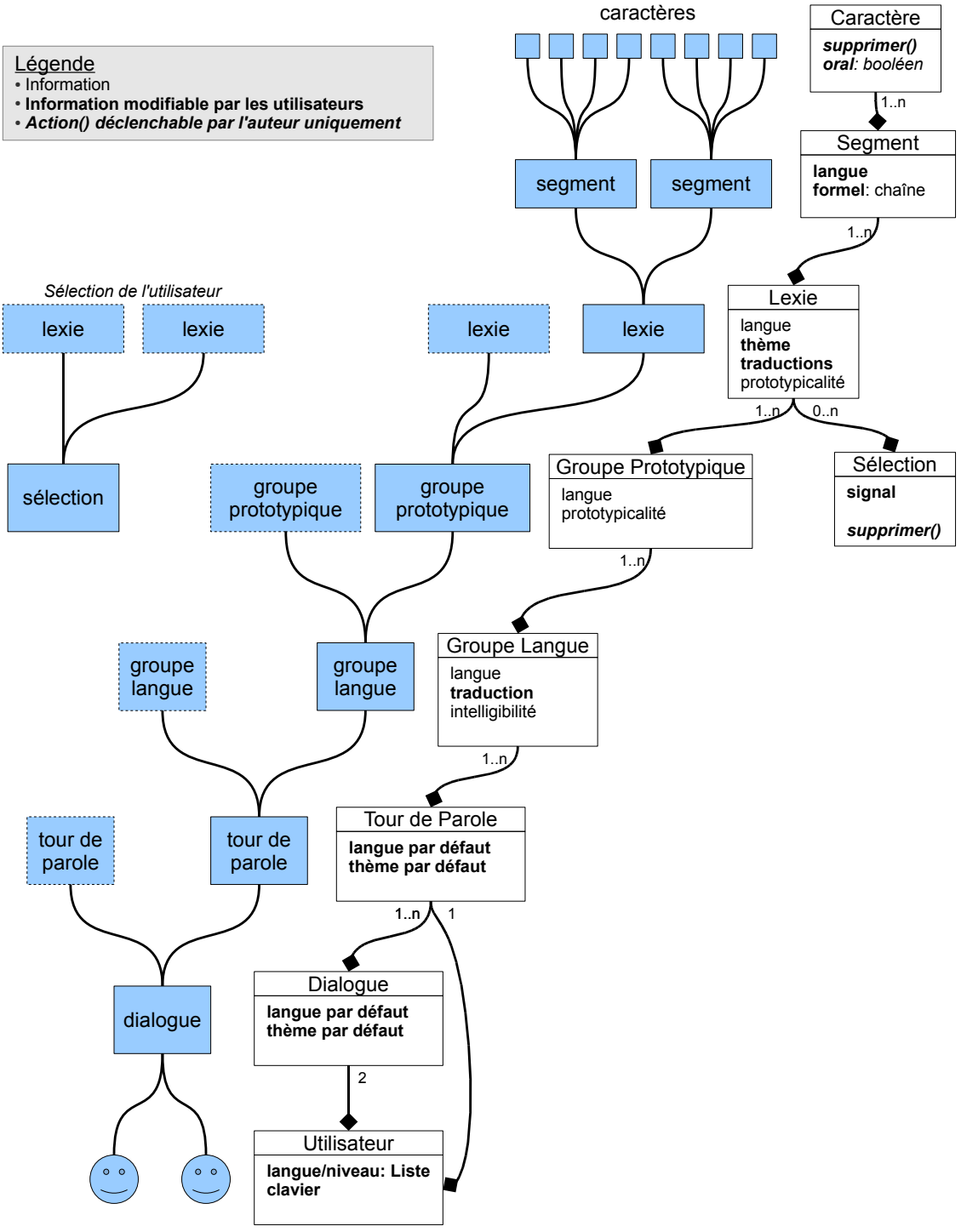


Figure 3.12 – Structure de données associée à un dialogue écrit

informations. Certaines de ces informations, par exemple la langue, sont pertinentes à plusieurs niveaux ; dans ce cas l'information de plus haut niveau est utilisée par défaut, et peut être redéfinie aux niveaux inférieurs. Chaque sous-structure (caractère, segment, mot, etc.) dispose de chaînages multiples : frère précédent, frère suivant,

liste des sous-structures de niveau inférieur, et le cas échéant structure de niveau supérieur.

Certaines informations (en gras sur la figure 3.12) peuvent être corrigées par les utilisateurs, et peuvent donner lieu à un sous-dialogue de clarification. Ces variables structurées sont de la forme : `(valeur calculée, {actions des utilisateurs})`, où `{actions des utilisateurs}` est la liste des actions effectuées par les utilisateurs par rapport à telle ou telle information associée à cette unité précise. Ces actions peuvent être :

Corriger : l'utilisateur corrige la valeur calculée automatiquement.

Signaler : l'utilisateur signale, à l'intention de son interlocuteur, que cette information comporte un problème.

Commenter : l'utilisateur ajoute un commentaire, ce qui entraîne automatiquement la création d'un signal. Ces commentaires peuvent s'enchaîner et forment alors un sous-dialogue de clarification.

Fermer : si un signal a été posé, l'utilisateur peut le retirer, ce qui a pour effet de clore l'éventuel sous-dialogue.

La valeur finale, « externe », d'une information, est celle déterminée en dernier par les utilisateurs, ou à défaut la valeur calculée automatiquement par le système.

Organisation des traitements

Comme pour Verbmobil, les modules sont indépendants et partagent une structure de données, dont ils peuvent être amenés à modifier la partie sur laquelle ils sont compétents (cf. figure 3.13). Certains traitements apportent des informations utiles pour les étapes antérieures (lignes pointillées), qui seront alors recalculées.

On se trouve alors confronté à un *problème d'interblocage* : deux traitements peuvent avoir besoin l'un de l'autre. Par exemple, le calcul de la langue bénéficie de la connaissance du thème, et réciproquement. Afin d'éviter ce genre de blocage, les calculs ne prennent en compte que les informations actuellement disponibles, et sont relancés dans un ordre déterminé. La figure 3.14 décrit l'organisation de ce genre de recalcul. En particulier, aucun traitement n'est lancé avant que ses prédécesseurs n'aient terminé leurs recalculs : dans l'exemple, le calcul $x+2$ n' est pas lancé avant l'obtention de la valeur définitive du calcul $x+1$.

En pratique, les modules de traitement sont simplement stockés dans une liste chaînée, parfois plusieurs fois (cas des recalculs), et appliqués séquentiellement.

Les interventions de l'utilisateur, lorsqu'il corrige une mesure, entraînent le recalcul de tous les successeurs (figure 3.15). Par le jeu des recalculs, les prédécesseurs peuvent aussi se retrouver impliqués (figure 3.16).

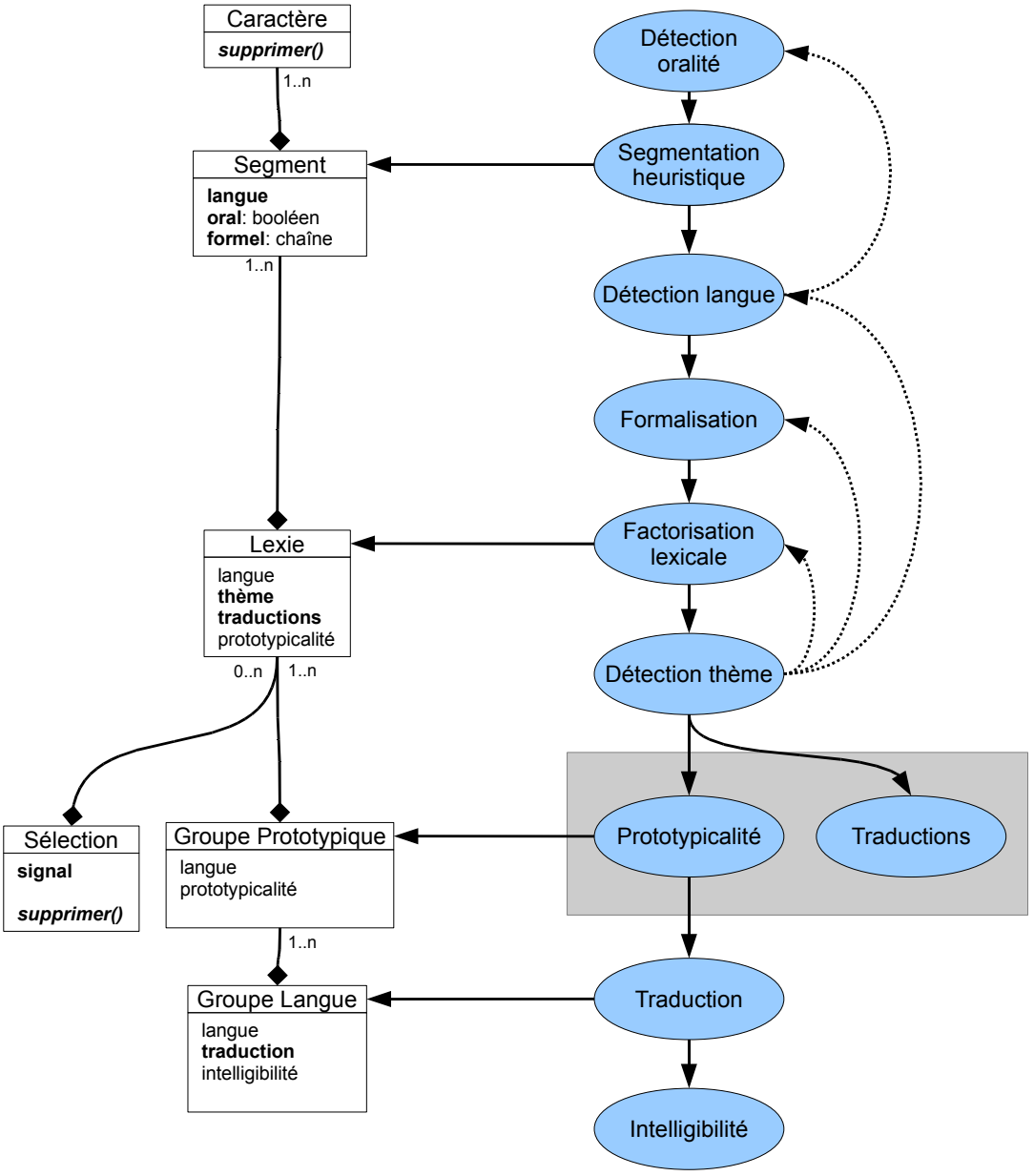


Figure 3.13 – Enchaînement séquentiel et rétroactif des traitements et création des unités.

En pratique, il s'agit simplement d'appliquer à nouveau tous les modules de traitement dans l'ordre de la liste, en commençant par la première occurrence du calcul corrigé par l'utilisateur.

5.2 Principes communs aux composants

5.2.1 Portée des mesures

Étant donné les niveaux de granularité progressifs des données, certaines informations peuvent être pertinentes à plusieurs niveaux. Dans ce cas, comme on l'a dit, l'information de plus haut niveau est utilisée par défaut, et peut être modifiée loca-

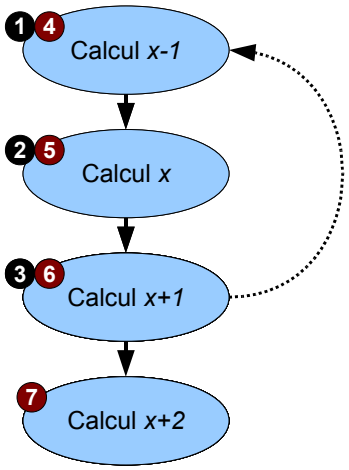


Figure 3.14 – Enchaînement des recalculs.

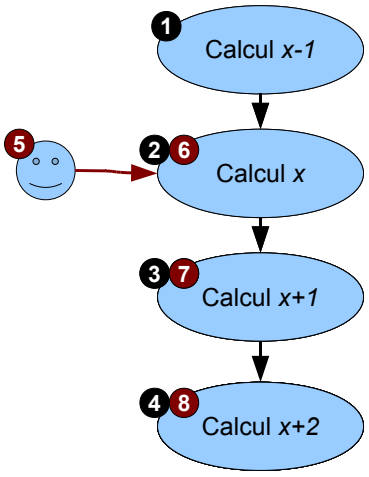


Figure 3.15 – Enchaînement des recalculs après une intervention de l'utilisateur.

lement. Par exemple, la langue est valable par défaut pour tout un tour de parole, mais au sein de ce tour peut se trouver un « groupe-langue » de langue différente, qui redéfinira la langue par défaut pour tous les membres du groupe.

Lorsque les informations portent sur un mot ou un groupe de mots, nous nous efforçons de prendre en considération son contexte. C'est à dire que ces informations ont en fait deux valeurs : une première locale, indépendante du contexte, et une seconde globale, « colorée » par les informations attachées aux mots de son contexte, pondérées en fonction de la distance. Ainsi le score contextuel s, pour une propriété donnée π, d'un mot m de rang x dans une suite de n mots (le contexte) est donné

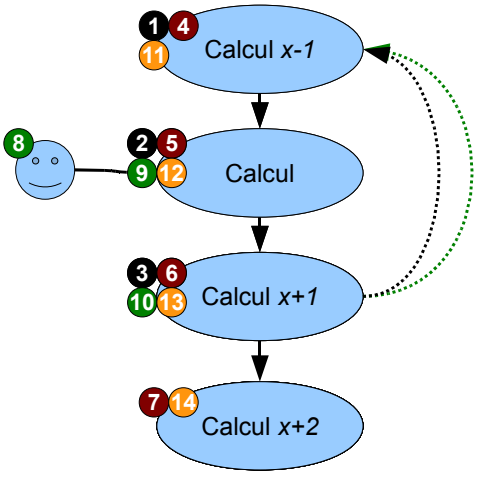


Figure 3.16 – Enchaînement des recalculs après une intervention de l'utilisateur, lorsqu'un recalcul est déjà prévu initialement.

par la formule suivante, où $dist()$ est la pénalité de distance :

$$\mathbf{s}_\pi(\mathbf{m_x}|\mathbf{m_1}\dots\mathbf{m_n}) = \frac{1}{n} \times \sum_{i=1}^{n} \frac{p_\pi(m_i)}{dist(m_x, m_i)} \tag{3.2}$$

Les tours de parole étant souvent courts en tchat, le contexte s'étend au dialogue tout entier. Cette pénalité de distance tient compte de la ponctuation : $np(x, i)$ donne le nombre de séquences de signes de ponctuation entre x et i. Elle tient aussi compte des limites de tours de parole : $ntp(x)$ donne le rang du tour de parole qui contient le mot d'indice x :

$$\mathbf{dist}(\mathbf{m_x}, \mathbf{m_i}) = \log(|(\mathbf{x} - \mathbf{i}) \times \mathbf{np}(\mathbf{x}, \mathbf{i}) \times (\mathbf{ntp}(\mathbf{x}) - \mathbf{ntp}(\mathbf{i}))| + \mathbf{1}) \tag{3.3}$$

À l'opposé, certaines informations associées à des groupes de mots seront déduites d'informations similaires de niveau inférieur, dont on a déjà calculé le score global. Pour $a > b$ (il faut donc au moins deux mots) :

$$\mathbf{s}'_\pi(\mathbf{m_a}\dots\mathbf{m_b}) = \frac{1}{b-a} \times \sum_{i=a}^{b} s_\pi(m_i|m_a\dots m_b) \tag{3.4}$$

Enfin, certaines informations n'ont de sens qu'au niveau du groupe, et dans le cas de la prototypicalité, servent même à constituer des groupes. Toujours dans le cas $a > b$:

$$\mathbf{s}''_\pi(\mathbf{m_a}\dots\mathbf{m_b}|\mathbf{m_1}\dots\mathbf{m_n}) = \frac{1}{b-a} \times \sum_{i=a}^{b} s_\pi(m_i|m_1\dots m_{i-1}, m_{i+1}\dots m_b) \tag{3.5}$$

5.2.2 Interface et préséance des mesures

Parfois, plusieurs valeurs peuvent partager un même mode de signalement, par exemple un soulignement. Si plusieurs de ces valeurs indiquent un problème, elles ne pourront pas être signalées en même temps. Dans ce cas, l'ordre de préséance est celui de la liste des traitements. Ainsi, les problèmes qui ont des conséquences les plus étendues pour le système sont traités en premier.

5.2.3 Interface et portée des sélections

Lors de son utilisation du système, l'utilisateur peut déplacer le curseur dans la séquence de caractères qu'il est en train de composer, ou bien sélectionner un fragment de la séquence. Il est important de savoir alors à quelle unité on doit associer les informations visualisées dans le menu contextuel. En l'occurrence, on considère l'unité de sélection comme étant l'unité lexicale détectée sous le curseur.

En ce qui concerne le positionnement du curseur, à partir du moment où il se situe sur une unité, c'est cette unité qui est considérée comme la sélection courante, et qui sera visible dans le menu contextuel. S'il se situe en dehors d'une unité lexicale (par exemple sur un signe de ponctuation), la sélection concerne le caractère.

Si une sélection intègre une unité lexicale ou un fragment d'unité lexicale, elle est automatiquement étendue jusqu'aux frontières lexicales. Sinon elle ne concerne qu'une séquence de caractères.

5.2.4 Incertitude

Dans un système où les traitements se suivent en cascade, les derniers modules sont les plus vulnérables aux erreurs commises en amont, et les erreurs auront tendance à s'accumuler. Le fait de soumettre à l'utilisateur les problèmes de bas niveau en priorité permet d'éviter de le distraire avec des problèmes en aval qui peuvent être issus d'erreurs en amont. Reste que si les erreurs s'accumulent en amont, les calculs effectués en aval perdent beaucoup de leur utilité.

C'est pourquoi, à chaque unité de traitement est associée un compteur de problèmes, basé sur le nombre de problèmes signalés à l'utilisateur sur cette unité. Au niveau supérieur, on fait la somme des compteurs des unités incluses ; si elle est supérieure à un seuil, le traitement s'arrête jusqu'à ce qu'il descende à une valeur acceptable. Ainsi certaines unités peuvent ne pas être définies, et certaines informations non disponibles. Les aides qui portent sur ces unités manquantes ou ces informations ne seront donc pas disponibles, jusqu'à la résolution des problèmes par l'utilisateur.

5.3 Composants passifs

5.3.1 Détection des marques de l'oralité

La première étape de traitement consiste à repérer les segments qui constituent des marques de l'oralité les plus évidentes, comme les émoticons et interjections. Pour les émoticons, on ne peut pas se contenter de laisser la segmentation se charger de les « nettoyer », puisque d'une part certains comportent des caractères alphanumériques, et d'autre part les outils de traduction automatique vont attendre un écrit formel, c'est à dire ponctué mais non « émotionné ».

Étant donné leur grande variabilité, on privilégie une recherche par dictionnaire. Toutefois, à cause de cette grande variabilité, les dictionnaires seront forcément incomplets, ce sera donc aux utilisateurs des les compléter, une tâche facilitée par le fait que, bien que ces marques puissent être très variées, chaque utilisateur se construit un lexique personnalisé très restreint.

Comme certaines de ces formes peuvent prendre des formes « allongées » (par exemple « bahhhhh »), le dictionnaire autorise l'utilisation des opérateurs classiques des expressions régulières.

Ces marques sont présentées à l'utilisateur sous forme de texte grisé (cf. figure 3.17), et peuvent être corrigées dans le menu contextuel à l'aide d'une case à cocher (cf. figure 3.18).

Figure 3.17 – Visualisation d'une marque de l'oralité.

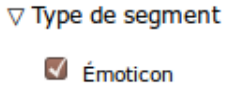


Figure 3.18 – Correction d'une marque de l'oralité.

5.3.2 Segmentation heuristique

La seconde étape de traitement consiste à segmenter la séquence de caractères. À cette étape, on ne fait qu'une seule hypothèse sur la langue, à savoir qu'elle est segmentée par des «blancs» à l'écrit. La méthode se veut donc agressive, générant un maximum de segments. La segmentation se base uniquement sur les « blancs » (classe POSIX `space`) et signes de ponctuation (classe POSIX `punct`). Les séquences numériques forment des tokens à part (par exemple « 50$ » formera deux tokens).

5.3.3 Détection de la langue

À l'écrit, le contexte et l'intelligence des locuteurs font que les problèmes d'identification de la langue sont rares ; c'est par contre très important pour un système de TALN. On pourrait certes se contenter de se baser sur la langue que les utilisateurs auront indiqué comme langue de conversation, mais comme on l'a vu, un tour de parole peut être multilingue. Chaque tour de parole a donc une langue par défaut, celle qu'a fourni l'utilisateur, que l'on peut estimer comme fiable la plupart du temps, et peut comporter des portions dans une ou plusieurs autres langues.

Le multilinguisme peut se manifester de deux manières différentes. Ce peut être un lexème « alloglosse » isolé : un terme de jargon (par exemple *booking* utilisé en français), un emprunt plus général, une citation d'un terme qui n'appartient pas au lexique de la langue de l'énoncé. Le changement de langue ne dépasse pas la portée du lexème, qui est intégré dans la syntaxe du reste de l'énoncé. À l'opposé, c'est tout un syntagme, ou même plus, qui peut être dans une autre langue, et obéir à sa syntaxe.

Nous avons donc besoin d'une identification de langue à deux niveaux : lexical et syntagmatique.

Actuellement, on utilise souvent pour détecter la langue des stratégies de classification automatique, basées sur des modèles n-grammes de caractères. Cette méthode est assez fiable, y compris en environnement bruité comme peut l'être un tchat, et permet même de détecter l'encodage des caractères [VT04]. Néanmoins, cela nécessite plusieurs dizaines de caractères pour être fiable. Cette méthode n'est donc utilisable que pour une recherche à l'échelle syntagmatique, et encore, à condition d'avoir des syntagmes de longueur suffisante ; le problème est qu'en tchat les tours sont assez brefs, on ne disposera donc même pas toujours de suffisamment de caractères pour détecter la langue d'un tour de manière fiable.

On peut donc envisager plutôt une recherche par dictionnaire, tenant compte du contexte pour les cas où la langue est ambigüe (par exemple, dans «Comment ça va ?», le mot *comment* est présent à la fois dans les dictionnaires français et anglais).

$$s_{lg}(m_x|m_1 \dots m_n) = \frac{1}{n} \times \sum_{i=1}^{n} \frac{p_{lg}(m_i)}{dist(m_x, m_i)} \tag{3.6}$$

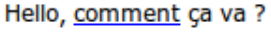


Figure 3.19 – Signalement d'un problème potentiel de détection de langue.

Figure 3.20 – Correction d'une détection de langue.

5.3.4 Formalisation

Même une fois les marques de l'oralité identifiées, d'autres spécificités du tchat peuvent poser problème, comme les abréviations, les allongements, et les variations phonético-graphiques. Le problème est que celles-ci, au contraire des marqueurs oraux, ne peuvent être purement et simplement « nettoyées », puisqu'il s'agit de « déformations » d'entités formelles que l'on doit conserver : en enlever une, c'est enlever un mot de l'énoncé.

En outre, les fautes d'orthographe, même triviales et dues à de simples erreurs de frappe, peuvent poser problème à deux niveaux. Le locuteur non natif aura plus de difficulté à reconnaître un mot ou la structure d'une phrase. Même une faute triviale a un coût cognitif pour le lecteur, qui dans le cas du locuteur non natif se trouve déjà impliqué dans une tâche cognitivement complexe. Au niveau du système, pour les autres composants d'aide, des mots non-identifiables car mal orthographiés posent bien évidemment un problème. Or, comme on l'a vu, les tchateurs sont plutôt soucieux de leur orthographe, du moins lorsqu'ils estiment qu'une faute peut nuire à la compréhension. Dans un contexte linguistique non natif, en présence d'outils nécessitant une grande rigueur orthographique, ce souci se trouve renforcé. Mais pour le locuteur non natif, encore plus qu'en tchat natif, ce sont les compétences qui peuvent faire défaut.

À cette étape nous traitons donc deux problèmes : les graphies informelles, volontaires, et les fautes involontaires.

On introduit donc une étape de formalisation, c'est à dire de transformation des segments informels en segments formels. Comme pour la détection de l'oralité, on se base sur un dictionnaire autorisant la recherche par expression régulière, mais qui cette fois associe à chaque segment informel un segment formel (un transducteur simple).

Les marqueurs de l'oralité repérés à l'étape précédente, lorsqu'ils ne sont pas associés à un symbole de ponctuation, sont ici marqués.

Toutefois, certains segments informels ne pourront pas être traduits, car sans équivalents formels, comme par exemple les émoticons.

Pour la correction des fautes involontaires, de nombreux outils de correction orthographique existent. Dans le cas d'une conversation en anglais, une telle correction orthographique « classique » peut suffire. Les énoncés sont segmentés automatiquement par « mot » (correspondant au niveau lexical dans notre système), chacun étant ensuite recherché dans un lexique de formes fléchies, ou bien préalablement lemmatisé avant d'être recherché dans un lexique de formes non fléchies (en particulier dans le cas de langues riches en compositions comme l'allemand et surtout pour les langues agglutinantes elles que le hongrois, le turc, etc.). En cas d'échec, le mot non trouvé est signalé (par exemple souligné) ; idéalement des mots du lexique, graphiquement proches, sont proposés en correction. Ce type d'aide ne concerne que les fautes dites « d'usage » (lexicales) et non celles dites « grammaticales » (morpho-syntaxiques). Plusieurs outils librement utilisables, tel qu'*Aspell* (projet *GNU*), implémentent ce type de système.

Dans le cas d'*Aspell*, le système renvoie une liste ordonnée de suggestions, classées par probabilité. Il est possible d'améliorer ce classement en tenant compte :

De la catégorie morphosyntaxique donnée par le lemmatiseur : les suggestions appartenant à la même catégorie sont privilégiées.

Du thème : selon la proximité thématique entre chaque suggestion et le thème calculé pour le contexte.

Le signalement des problèmes et leur mode de correction sont assez similaires à ceux de la détection de la langue.

Figure 3.21 – Signalement d'un problème d'orthographe (ou de mot absent du dictionnaire).

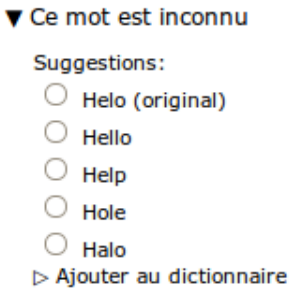


Figure 3.22 – Correction d'un problème d'orthographe.

La correction orthographique aura en outre l'intérêt de signaler certains barbarismes, même si les corrections suggérées, basées uniquement sur la forme, risquent de ne pas être en elles-même d'une grande utilité dans ce cas. Il reste néanmoins utile de les signaler à l'utilisateur.

5.3.5 Lemmatisation

Pour les traitements lexicaux ultérieurs, nous aurons besoin d'associer à chaque item lexical son lemme. Nous utilisons donc un lemmatiseur, *Treetagger* [Sch94], développé par Helmut Schmid de l'université de Stuttgart. Il s'agit d'un analyseur statistique, qui tient compte du contexte en cas de doute sur un lemme. Grâce à de nombreuses contributions, ce lemmatiseur est capable de traiter 11 langues modernes : allemand, anglais, bulgare, chinois, espagnol, français, grec, italien, néerlandais, portugais et russe. Il est facilement extensible à d'autres langues, sachant néanmoins qu'il se base sur des dictionnaires de suffixes et de préfixes (ces derniers ont été ajoutés pour le traitement de l'allemand [Sch02]), et semble donc peu adapté en l'état à des langues agglutinantes (hongrois, turc, finnois...) risque de poser problème. L'utilisateur peut corriger le lemme, de la même façon que pour les autres outils.

Ce lemmatiseur associe à chaque forme son lemme et sa catégorie syntaxique (partie du discours). Pour l'anglais, les catégories utilisées sont celles du *Penntree Bank* [San90], un standard reconnu. Une information minimale sur la flexion est parfois disponible, par exemple pour les verbes (infinitif, participe passé, etc.). Pour l'instant, nous ne nous servons pas de ces dernières informations. Toutefois, en cas d'intégration avec des outils sensibles à la syntaxe, ou d'une intégration plus étroite de la traduction automatique, ces informations pourraient s'avérer utiles, car elles permettraient de détecter des erreurs potentielles et de demander confirmation des hypothèses morphologiques du système.

Treetagger ne traite pas les mots composés, ainsi « pomme de terre » sera analysé comme une séquence de trois lemmes. Pour identifier ce type de lexème, nous avons besoin d'une étape de « décomposition » lexicale.

5.3.6 Factorisation lexicale

Les lemmes des segments sont ensuite recherchés dans à un dictionnaire de formes composées, trié du plus grand nombre de segments au plus petit. Cela permet de repérer tous les mots composés ou expression lexicalisées, en favorisant les lexèmes les plus longs. Cette stratégie est largement utilisée pour les langues asiatiques sans segmentation lexicale, et donne de bons résultats. On obtient alors l'une des unités les plus importantes du système, le lexème.

L'utilisateur peut corriger le système, soit en annulant une factorisation erronée (cf. figure 3.23), soit en demandant la factorisation en un seul item lexical d'une sélection de plusieurs items lexicaux.

Par contre, le fait de signaler les ambiguïtés de segmentation ne nous a pas paru souhaitable, car cela reviendrait en fait à signaler quasiment toutes les mots composés comme ambiguïtés potentielles, ce qui serait assez lourd pour l'utilisateur. En

Figure 3.23 – Correction d'une factorisation.

effet, dans le cadre de cet algorithme et des données disponibles (rarement lexicalisées manuellement), il n'est pas possible de mesurer réellement l'ambiguïté segmentale.

5.3.7 Détection du thème

Le thème donné par l'utilisateur n'étant pas une information totalement fiable, nous préférons la considérer comme une initialisation facultative d'un paramètre qui évoluera automatiquement au fil du dialogue. Le thème, initialisé ou non, devrait donc ensuite être adapté aux productions des locuteurs. Afin de détecter automatiquement le thème de tours de parole, nous nous basons sur la méthode mise au point par Dominique Vaufreydaz pour la détection automatique de thèmes à partir des *newsgroups* [Vau02].

Le principe est simple : à partir de chaque groupe de discussion, un modèle de langage unigramme thématique est extrait. Les groupes étant hiérarchisés (par exemple *fr.petites-annonces.informatique.logiciel*), on structure les modèles de la même manière, incrémentalement (par exemple, pour reprendre l'exemple précédent, on aurait un modèle générique *fr*, un autre un peu plus spécifique *fr.petites-annonces*, puis *fr.petites-annonces.informatique* encore plus spécifique).

Cette méthode affecte un poids à chaque mot de la séquence recherchée, inversement proportionnel à sa fréquence dans l'ensemble des thèmes. Cela permet d'écarter le bruit causé par les mots courants quel que soit le thème, comme les mots-outils (articles, prépositions, etc.). Ainsi, chaque mot de la séquence S dont on veut déterminer le thème est recherché dans le modèle testé, et la probabilité *p(S)* que cette séquence appartienne à ce thème est égale au produit de la probabilité p_θ de chaque mot dans le thème considéré, compte tenu de sa probabilité p_Ω dans le reste du corpus (équation 3.7).

$$p_\theta(S) = \prod_{i=1}^{n} \frac{p_\theta(m_i)}{p_\Omega(m_i)} \tag{3.7}$$

Dans notre système, la pondération par pertinence (division par $p_\Omega(m_i)$) est incluse dans la formule de calcul du score s_π (équations 3.2 et suivantes). La formule doit donc être simplifiée pour être incluse dans le système :

$$p_\theta(S) = \prod_{i=1}^{n} p_\theta(m_i) \tag{3.8}$$

À l'aide cette méthode, on obtient le score de la séquence pour chaque thème de l'arborescence. On peut ensuite soit prendre le thème de score le plus élevé, soit plusieurs thèmes dominants. Dans ce dernier cas, on considère en fait le thème comme un paramètre complexe, constitué d'une arborescence de paires thème-probabilité.

L'espace de recherche arborescent peut être réduit dynamiquement par élagage. En parcourant l'arborescence en largeur d'abord lors de la recherche, on peut décider d'abandonner l'évaluation des descendants des nœuds les moins prometteurs (approche *n-best*).

Si les utilisateurs ont spécifié un thème pour le dialogue, les probabilités calculées peuvent être pondérées par ce choix. On part du principe que le thème est spécifié par l'utilisateur de façon absolue, par exemple à l'aide de cases à cocher. La probabilité annoncée par l'utilisateur p_{util} peut alors prendre deux valeurs, déterminées en fonction du degré de confiance envers l'utilisateur, mais aussi du fait que pour ce dernier, ne pas cocher un thème dans l'interface ne signifie pas qu'il interdit le thème : par exemple 0,5 et 1. L'équation 3.7 peut alors être complétée comme suit :

$$\mathbf{p}_\theta(\mathbf{S}) = p_{util} \times \prod_{i=1}^{n} p_\theta(m_i) \tag{3.9}$$

5.3.8 Ambiguïtés

Les ambiguïtés linguistiques ont, pour le français, été décrites en détail par Catherine Fuchs [Fuc00]. Pour le TALN, une description formelle a été élaborée [BBL04] [BT95].

Dans le cadre de nos aides, nous nous en tiendrons à une définition *ad hoc* plus restrictive, réduite à un sous-ensemble de la classe des ambiguïtés morpho-lexicales. Nous parlerons d'ambiguïté lorsque dans notre dictionnaire bilingue, un mot m a plusieurs traductions $\mu^0 \dots \mu^n$. Par exemple *bank* peut se traduire : « banque ; bord, rive ».

Certaines ambiguïtés sont plus importantes que d'autres. Ainsi, *bank* :: *remblai* ne diffère guère de *bank* :: *talus*, et confondre ces deux acceptions aurait moins de conséquence qu'avec *bank* :: *banque*. Autre exemple : le russe ne distingue pas le bras et la main. Dans cette langue, c'est un tout, appelé *рука*. Ce terme est donc ambigu dans sa traduction vers le français, mais une erreur n'est pas nécessairement problématique pour la compréhension, les deux termes n'étant pas très éloignés sémantiquement.

Pour pouvoir hiérarchiser les ambiguïtés, nous avons besoin de pouvoir calculer la distance sémantique en langue cible entre les traductions. En première approche, nous nous basons sur une mesure simple calculée à partir de notre modèle de langage trigramme : la distance paradigmatique.

5.3.9 Distance paradigmatique

La distance paradigmatique entre deux termes désigne leur probabilité de partager le même contexte. Ce calcul s'inspire de celui des voisins distributionnels développé par [BL02], mais sans analyse syntaxique préalable, et se révèle donc moins précis. Ainsi, si a et b sont deux termes, $co(a, b)$ le nombre de leurs contextes communs

et $ne(a, b)$ le nombre de leurs contextes différents, nous calculons un coefficient de Jaccard :

$$\mathbf{para}(\mathbf{a}, \mathbf{b}) = \frac{co(a, b)}{ne(a, b)} \tag{3.10}$$

À partir de cette mesure, on peut calculer **rpara**$(\mathbf{m}, \mathbf{n})$ une fonction qui renvoie les n premiers voisins paradigmatiques de m.

Nous construisons notre modèle de langage paradigmatique à partir de cette mesure.

5.3.10 Distance syntagmatique

En complément, on utilise aussi la distance syntagmatique **synta**$(\mathbf{a}, \mathbf{b})$. Elle sert à déterminer si deux expressions apparaissent fréquemment l'une à proximité de l'autre, et ont donc, on le suppose, une relation d'ordre thématique.

Nous construisons notre modèle de langage syntagmatique à partir de cette mesure. À la différence d'un modèle n-grammes, il n'est pas sensible à l'ordre des mots, mais au fait qu'ils apparaissent fréquemment à proximité l'un de l'autre.

5.3.11 Calcul de la traduction lexicale

Nous partons d'un énoncé comportant plusieurs ambiguïtés (figure 3.24), c'est à dire que certains mots m_i ont plusieurs traductions μ_i^n. Comme on souhaite se baser

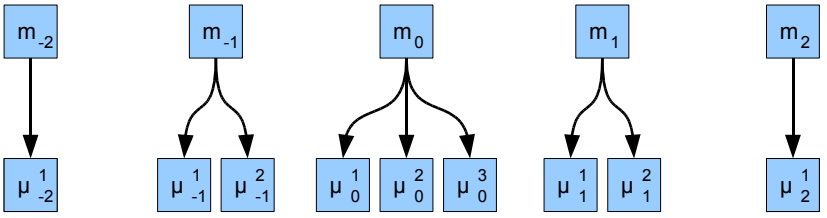


Figure 3.24 – Dans une séquence de cinq mots m_i, plusieurs traductions μ_j^i pour chaque mot.

sur le contexte pour calculer le score des traductions, nous adoptons une structure en treillis (figure 3.25). Trouver la meilleure traduction pour chaque mot revient alors à trouver le meilleur chemin à travers ce treillis. Comment calculer le score d'un tel chemin ? Nous nous basons sur notre modèle syntagmatique pour retrouver les traductions qui, en langue cible, apparaissent fréquemment ensemble.

Nous pouvons ainsi déterminer automatiquement une traduction en cas d'ambiguïté. On peut, par la suite, signaler à l'utilisateur les cas où le résolution automatique de l'ambiguïté n'est pas satisfaisante, c'est à dire quand aucune traduction

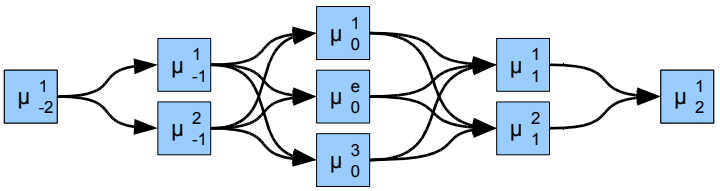


Figure 3.25 – Graphe des chemins possibles.

n'émerge du lot, alors que ces traductions ont des significations très différentes (distance paradigmatique élevée). Cela peut en effet être une source de confusion pour les utilisateurs.

5.3.12 Expression de l'ambiguïté lexicale

Pour que l'utilisateur puisse intervenir sur le résultat du calcul de la traduction lexicale, il faut pouvoir les lui présenter. Pour le locuteur non natif, il suffit de lui demander de choisir dans la séquence μ la traduction adéquate.

Figure 3.26 – Correction d'une ambiguïté lexicale par un locuteur non natif.

Par contre, les choses se compliquent si l'on souhaite présenter l'ambiguïté à un locuteur natif . Par exemple, l'ensemble de traductions « rive ; bord ; remblai ; talus ; berge ; banque » se rétrotraduit par *bank, border, edge, shore ; railroad embankment ; bank ; bank ; bank*. Ce problème est représenté par la figure 3.27. Dans ce cas, on utilise

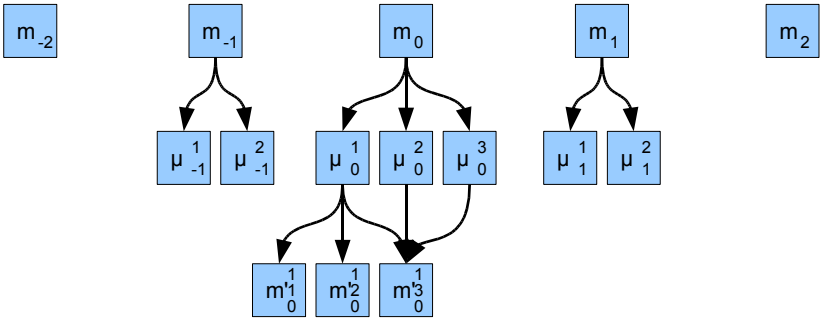


Figure 3.27 – Rétrotraductions (m'). Seules les deux premières sont désambiguïsantes.

des rétrotraductions désambiguïsantes lorsque c'est possible (ici, pour « rive » : *border, edge, shore*), ou bien une traduction de synonymes obtenus à partir d'un dictionnaire monolingue. Lorsque cette dernière ressource n'est pas disponible, nous

recherchons un contexte désambiguïsant dans notre corpus, c'est à dire un énoncé dans lequel la probabilité de la traduction que l'on cherche à exposer est la plus forte. Un exemple est donné en figure 3.28 pour la traduction de *bank*.

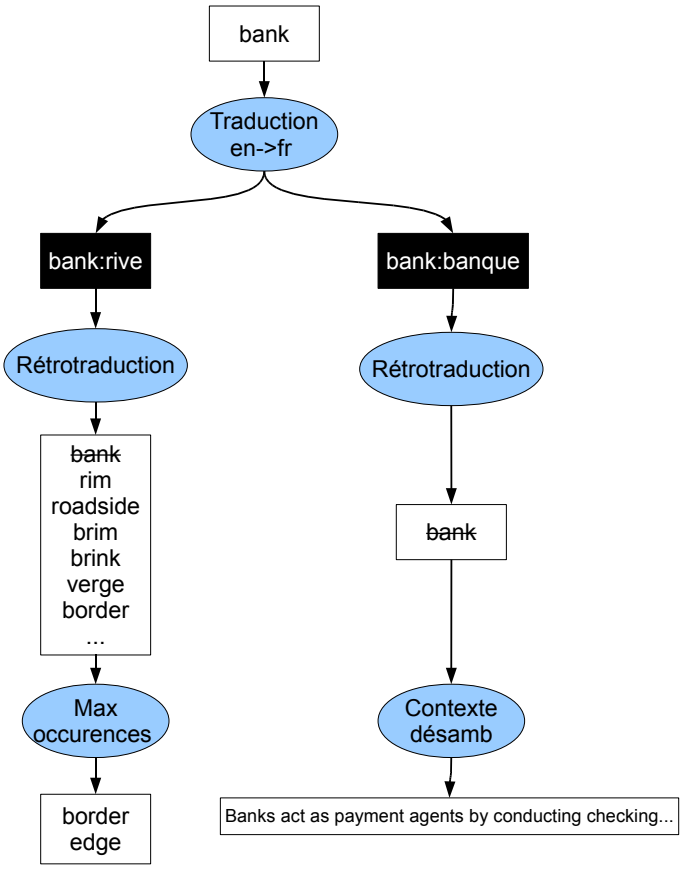


Figure 3.28 – Calcul de l'expression en anglais de l'ambiguïté de la traduction du mot *bank* en français.

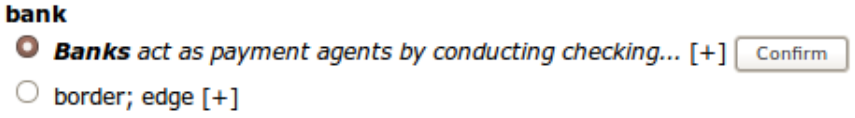


Figure 3.29 – Correction d'une ambiguïté lexicale par un locuteur natif.

5.3.13 Mesure de la prototypicalité

Un problème important pour les locuteurs non natifs, en particulier face à un locuteur natif, est de savoir si les expressions qu'ils utilisent sont idiomatiques, c'est à dire si elles sont réellement attestées dans la langue. D'autre part, cela peut contribuer à la segmentation linguistique des énoncés, lorsque les frontières linguistiques sont brouillées par des termes appartenant à plusieurs langues.

Selon la théorie du prototype, les éléments les plus centraux d'une catégorie servent de point de référence cognitif [Ros73] [Ros75]. Par exemple, un pigeon ou un rouge-gorge sont « plus » des oiseaux qu'un manchot ou une autruche, ou encore une chaise est « plus » un meuble qu'un téléphone. Ces perceptions subjectives sont validées sur le plan expérimental : par exemple, les sujets catégorisent plus vite les membres prototypiques que les autres. L'importance de la prototypicalité est envisagée dans le cadre de l'acquisition du langage [Slo81] [WKL96], la grammaire étant déduite par l'enfant à partir de modèles concrets (de prototypes), et chez l'adulte comme traitement cognitif pré-syntaxique, les énoncés étant d'abord comparés à des modèles concrets avant le passage au niveau syntaxique [Bre06].

Dans un énoncé, certaines formulations paraissent plus « normales » que d'autres. Comme on l'a vu en première partie, certaines de ces formulations sont conventionnelles et attendues [DS95] (réponses contraintes, scripts), et jouent un rôle à l'oral pour le décodage acoustico-phonétique des énoncés, en contraignant considérablement l'espace des énoncés possibles. Ce rôle existe aussi à l'écrit, où les mots et expressions ne sont pas décodés caractère par caractère mais reconnus grâce à leur forme globale et à leur contexte, d'où parfois des erreurs de décodage graphique nécessitant une relecture.

Il est donc utile d'évaluer la prototypicalité d'un énoncé, et plus précisément la répartition de la prototypicalité dans un énoncé : quelle partie est prototypique et quelle autre ne l'est pas ? En première approche, un simple modèle de langage n-grammes peut faire l'affaire.

5.3.14 Groupe linguistique et traduction

Comme on l'a vu, la langue doit aussi être calculée au plan syntagmatique. On ne peut pas se contenter d'un changement de langue sur le plan lexical pour former les groupes linguistiques : ce changement, même confirmé par l'utilisateur, n'est pas forcément significatif du point de vue de la syntaxe.

C'est pourquoi l'on se base plutôt sur les groupes prototypiques calculés précédemment, qui donnent un degré de prototypicalité appliqué à une langue donnée. Une séquence de groupes prototypiques de même langue forme alors un groupe linguistique.

5.3.15 Mesure de l'intelligibilité

Pour un locuteur natif qui s'adresse à un non natif, mais aussi pour ce dernier lorsqu'il se lance dans des constructions syntaxiques complexes, il est utile de veiller à l'intelligibilité des propos.

Il a été proposé une mesure de l'intelligibilité qui nous parait relativement adaptée, et qui consiste simplement à comparer les scores BLEU du texte original et de sa

rétrotraduction [IUI06]. Ce score, appliqué à un groupe linguistique, permet d'en estimer l'intelligibilité.

Lorsque l'intelligibilité est basse, cela est signalé en regard du tour de parole.

Il faut néanmoins tenir compte du fait que les mesures automatiques telles que BLEU sont assez souvent critiquées [Bla04b]. Si l'idée de comparer texte original et rétrotraduction nous semble intéressante, le problème d'une « bonne » évaluation des traductions reste entier.

5.3.16 Traduction automatique vers la langue de l'utilisateur

L'utilisateur peut consulter une traduction automatique du message dans sa langue.

5.3.17 Sous-titrage lexical

Cet outil est inspiré par une idée de Christian Boitet pour l'assistance à la compréhension (orale) des langues étrangères [Boi98], et adapté ici pour l'écrit. Il s'agit de « sous-titrer » le message au niveau lexical (c'est à dire qu'on ne traduit pas les énoncés, mais chaque mot isolément), mais uniquement les mots difficiles en fonction du niveau de l'utilisateur. La « difficulté » d'un mot est calculée en fonction de son nombre d'occurrences dans le corpus associé au thème en cours.

Le locuteur natif n'a pas besoin de sous-titrage pour comprendre le texte, mais cela peut l'alerter sur des erreurs commises par le système. C'est pourquoi, lorsque la rétrotraduction lexicale donne des synonymes désambiguïsants, ceux-ci peuvent servir de sous-titres.

5.4 Composants actifs

5.4.1 Post-édition

Tous les messages sont modifiables par leur auteur. Comme on l'a vu, les tchateurs corrigent fréquemment leurs messages *a posteriori*, bien que les logiciels de tchat ne soient pas prévus pour cela.

Les versions antérieures à chaque modification sont sauvegardées dans un historique. Les deux locuteurs peuvent le consulter, mais seul l'auteur du message peut revenir à un état antérieur.

5.4.2 Dictionnaire

Des dictionnaires bilingues « classiques » sont accessibles par les utilisateurs.

5.4.3 Livre de phrases

Des livres de phrases (simples et à variables lexicales) sont accessibles, en fonction du thème de la conversation.

5.4.4 Traduction automatique vers la langue de la conversation

L'utilisateur peut composer une partie de message dans sa langue, puis la sélectionner pour la traduire dans la langue de la conversation.

5.4.5 Signal

Les utilisateurs peuvent attacher un signal aux unités (segment, mot, etc.). Il existe un signal prédéfini « je ne comprends pas », mais les utilisateurs peuvent attacher le texte de leur choix.

De plus, chaque paire unité/module peut se voir attacher un signal. Ainsi, on peut signaler précisément un problème de désambiguïsation automatique, de détection de langue, etc.

Ce signal peut être retiré à tout moment.

5.4.6 Sous-dialogue de clarification

L'interlocuteur peut réagir à ce signal, en modifiant son message (reformulation), en corrigeant le système, ou bien en répondant au message : il lance alors un sous-dialogue de clarification.

5.4.7 Éléments « chauds »

Une unité qui reçoit un signal devient « chaude », c'est à dire qu'elle est mise en évidence par rapport au reste du texte. Cet état disparait lorsque le signal est retiré.

6 Outils graphiques

Les outils graphiques viennent en complément des outils de dialogue, et permettent aux utilisateurs d'échanger des documents (pages web, images, plans, etc.), et d'intervenir dessus collaborativement. Ainsi, chaque utilisateur peut dessiner par-dessus ces documents à la souris (chacun dispose d'une couleur qui lui est propre). Il peut aussi déposer des signaux sur des points difficiles (par exemple passage délicat sur un plan), qui peuvent donner lieu à un sous-dialogue écrit. En fonction de l'élément sous-jacent, l'utilisation de l'outil peut varier :

Tableau blanc : aucun élément sous-jacent.

Image : l'élément sous-jacent est une image, par exemple un plan.

Page Web : l'élément sous-jacent peut être n'importe quelle page Web.

Encyclopédie : l'élément sous-jacent est une page de Wikipédia.

Dans les deux derniers cas, le mode « dessin » peut être désactivé pour permettre la navigation, mais par un seul utilisateur à la fois. Les changements d'URL et le défilement sont synchronisés.

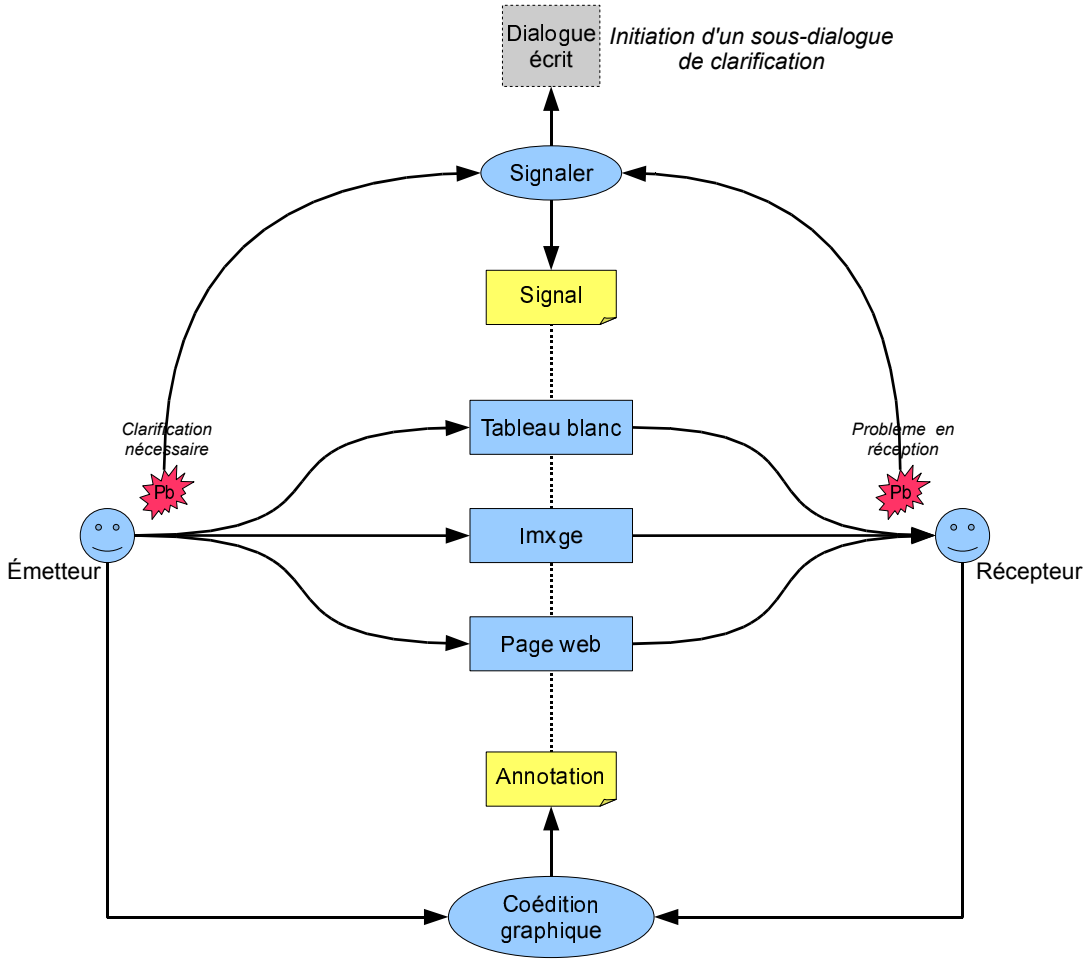


Figure 3.30 – Organisation des aides du mode graphique.

Conclusion

Nous avons vu dans cette partie comment aider les locuteurs dans le cadre d'un dialogue spontané, où au moins l'un des participants est non natif.

Il ne s'agit pas tant de créer de « meilleurs » composants, ce qui étant donné le nombre d'aspects à traiter, serait de toute façon impossible dans le temps d'une thèse, mais d'organiser des composants classiques de façon à simplifier le travail métalinguistique collaboratif que les locuteurs essayent déjà de faire sans aide. Dans cette optique, les composants peuvent même être mauvais, l'important est de leur intégrer une dimension interactive et collaborative permettant aux utilisateurs de dépasser largement les limitations et les erreurs des composants.

La partie suivante présentera quelques problèmes liés à l'implémentation de ce système.

Chapitre 4

Implémentation

1 Krater : bibliothèque de prototypage d'applications Web

Dans un contexte d'utilisation ponctuelle, afin de débloquer des « dialogues d'affaires », la nécessité d'installer un logiciel peut être un obstacle important à l'utilisation de notre outil, surtout dans un contexte professionnel où l'utilisateur ne peut pas toujours installer simplement un logiciel. De plus, les proxys d'entreprise peuvent bloquer tout ce qui n'est pas du « Web » (c'est à dire protocole HTTP sur port 80).

C'est pourquoi nous avons opté pour une implémentation en tant que site Web dynamique. En 2006, lorsque nous avons commencé l'implémentation, ce domaine était en pleine effervescence et de nombreux outils ouverts se proposaient d'implémenter ce genre de site. Aucun outil de référence n'avait encore émergé et les outils n'étaient pas toujours mûrs ; nous avons été tributaire de ce contexte bouillonnant.

L'implémentation sous forme d'application web soulève des problèmes génériques n'ayant rien à voir avec la problématique de l'assistance au dialogue en langue seconde (interface dynamique, synchronisation entre les utilisateurs, etc.), c'est pourquoi nous avons développé une bibliothèque indépendante pour les prendre en charge, indépendamment de l'aspect d'aide au dialogue. Nous avons appelé *Krater* cette bibliothèque destinée à être utilisée par notre outil d'aide *Koinè*.

1.1 Krater-1 : l'extension Firefox intégrée

Dans un premier temps, nous avons tenté de centrer le problème sur la synchronisation des données entre utilisateurs (cf. figure 4.1). Une conversation donc était considérée comme une donnée à synchroniser entre les utilisateurs, et dont dérivait l'interface.

Afin d'éviter les problèmes de corruption de données en cas de modifications concurrentes, chaque donnée partagée pouvait avoir autant de valeurs internes que

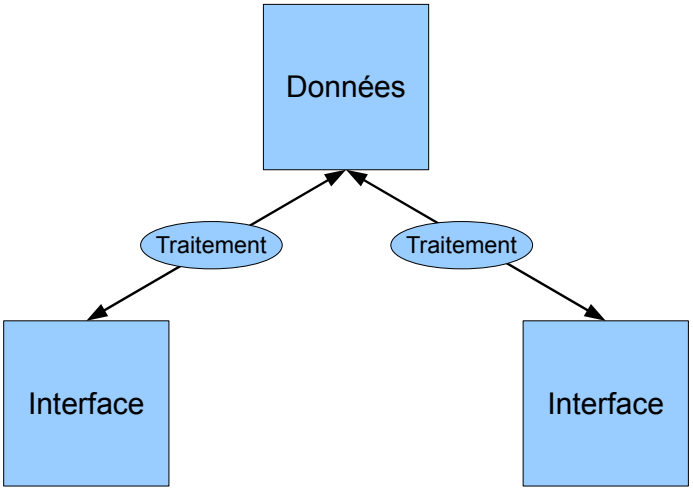


Figure 4.1 – Principe de Krater-1.

d'utilisateurs. Chacune de ces valeurs était datée. La valeur externe était celle de la valeur interne la plus récente. Cette stratégie d'unification des versions a été conservée dans toutes les versions ultérieures.

Le navigateur Firefox présentait l'environnement idéal pour ce type de développement. En effet, à côté de la plate-forme XPFE[1] dédiée à la conception d'interfaces riches, était intégrée la librairie XPCOM[2], permettant de construire automatiquement, selon des règles au format XBL[3], une interface XUL[4] à partir de données RDF[5].

Les composants de traitement des messages étaient développés en tant que services Web autonomes, sous forme de servlets Java, et intervenaient sur les données partagées sur le mode du tableau noir. Cette architecture avait l'intérêt de suivre une structure modèle-vue-contrôleur rigoureuse. Mais surtout, grâce à l'intégration de XBL au cœur de la plate-forme, les rafraîchissements de l'interface étaient transparents pour l'utilisateur : les données modifiées étaient répercutées dans l'interface, sans perte de focus ni « clignotement » désagréable.

[1] *Cross Plaform Front-End* : maintenu par la fondation Mozilla et permettant le développement d'applications locales et en ligne à l'aide des langages du Web (XUL, XHTML, CSS, JavaScript). Il sert de base à des logiciels aussi divers que le navigateur Firefox, le client courriel Thunderbird, l'éditeur Web Nvu, le calendrier Sunbird, ou encore le jukebox Songbird.

[2] *Cross-Platform Component Object Model.*

[3] *XML Bindings Language* : format de description de composants créé par Mozilla.

[4] *XML-based User interface Language* : langage de description d'interfaces assez semblable à XHTML, mais plus riche.

[5] *Resource Description Framework* : format de graphe développé par le W3C, utilisé notamment pour le codage sémantique (ontologies, Web sémantique).

Toutefois, au fil du développement, le format RDF s'est avéré particulièrement compliqué à maintenir. En effet, si ce format est bien adapté à la description de graphes sémantiques complexes, il ne l'est pas vraiment pour de simples paires attribut-valeur. Mais surtout, la définition de règles par le langage XBL est très contrainte, et certains concepts simples sont difficiles à traduire en XBL, comme par exemple les sauts conditionnels. La conception de l'interface devint donc rapidement laborieuse, et la moindre retouche nécessitait des efforts importants.

1.2 Krater-2 : utilisation de langages plus standard

Nous avons donc développé une seconde version, reprenant les mêmes principes et une partie des composants, mais en remplaçant les règles XBL par une combinaison de XSLT (toujours des règles donc) et de JavaScript, manipulant à l'aide du DOM les données (alors au format XML et non plus RDF) avant le traitement, puis l'interface produite après le traitement.

L'ajout d'un prétraitement et d'un post-traitement procéduraux permirent de dépasser les limites des transformations déclaratives, même si la gestion de ces trois étapes restait un peu complexe. Mais le passage de Firefox à la version 3 montra les limites d'un outil massivement basé sur le JavaScript, la fondation Mozilla ayant décidé de modifier l'implémentation du DOM, ce qui rendait l'application inutilisable sur Firefox 3. Il en découlait que probablement à chaque version de Firefox, il y aurait un travail de maintenance à effectuer, ce qui était difficilement acceptable pour une application qui, déjà, se limitait au seul moteur de rendu Gecko[6].

1.3 Krater-3 : le formulaire riche

Une troisième version fut donc réalisée. Elle visait à réimplémenter l'ensemble des fonctionnalités grâce au langage XFORMS, mis au point par le W3C, et permettant la mise en forme dynamique d'interfaces XHTML, à partir des données au format XML. En outre, l'interface est totalement intégrée avec les données : toute modification d'un élément de l'interface lié au modèle est aussitôt répercutée, sans qu'il soit nécessaire de coder quoi que ce soit.

Trois implémentations de XFORMS libres et réellement supportées par une communauté étaient disponibles :

Firefox : sous forme d'extension. Mais seul un sous-ensemble de XFORMS est effectivement implémenté.

Orbeon : XFORMS est totalement intégré dans ce gestionnaire de contenu. Mais cette intégration rend très difficile de faire autre chose que ce que ce pourquoi le gestionnaire a été à conçu, à savoir des sites Web « classiques ».

Chiba : prise en charge complète d'XFORMS, sans contrainte d'utilisation.

[6]Moteur de rendu de Firefox, mais aussi des applications basées sur XPFE et des navigateurs « compatibles Firefox » tels que Camino (MacOS X) et Epiphany (Gnome).

Le principal problème est que ces trois implémentations se révèlent assez largement boguées, dès lors que l'on sort des sentiers battus. Arriver simplement à afficher une donnée vectorielle sous forme de liste, en grisant une ligne sur deux pour faciliter la lecture, relève du parcours du combattant.

1.4 Krater-4 : retour à des choses simples

Finalement, quelque peu lassé de ces technologies présentées comme révolutionnaires, et aux fonctionnalités réellement intéressantes, mais à la mise en œuvre complexe et limitante, nous nous sommes résolus à une implémentation complète d'une bibliothèque pour le développement de prototypes d'applications Web simples, à partir d'outils éprouvés, et facile à mettre en œuvre, suivant en cela la philosophie KISS[7] ; à savoir le langage PHP et une interface en XHTML.

Concrètement, un développement Web classique permet déjà de convertir (flèches noires dans la figure 4.2) des données de toutes sortes (base de données relationnelle, divers formats de fichiers, divers types de variables) en interface HTML, à l'aide de langages impératifs évolués (PHP, JSP, ASP, Perl, Ruby, etc.). Le problème de la synchronisation se règle simplement en centralisant toutes les données sur le serveur, tout en conservant la stratégie d'unification de versions développée pour Krater-1. Restent deux problèmes :

- la rétropropagation des modifications de l'utilisateur sur l'interface vers les données qui l'ont générée, pour lesquels il n'existe pas vraiment de solution simple (flèches grises dans la figure 4.2).
- la mise à jour dynamique de l'interface lorsqu'elle est regénérée.

1.4.1 Rétropropagation des modifications de l'utilisateur

En HTML, il existe un composant dédié à l'entrée de données par l'utilisateur, il s'agit des formulaires. Ceux-ci contiennent des champs, qui peuvent être de diverses formes (zone de saisie, case à cocher, invisibles, etc.) ; ces formes sont assez limitées, mais toute intervention de l'utilisateur sur un élément graphique peut être associé à un champ invisible, grâce à des événements JavaScript (*onclick*, *onmouseover*, *onkeypress*, etc.). On peut donc associer sans difficulté tout élément à un champ nommé. Les formulaires peuvent être envoyés automatiquement à la moindre modification[8], leur contenu est alors envoyé au serveur et peut être récupéré sous forme de tableau associatif, et le code de génération de la page peut se baser dessus pour regénérer la page, en tenant ainsi compte des actions de l'utilisateur.

Le principe est alors de créer un type de variable, dont les noms correspondent aux champs des formulaires, et la valeur aux valeurs entrées par l'utilisateur (soit directement par le biais d'un champ visible, soit par le truchement d'un événement

[7] *Keep It Simple, Stupid*

[8] Cela peut-être configuré ; pour les champs de saisie, le comportement par défaut est d'attendre la perte du focus.

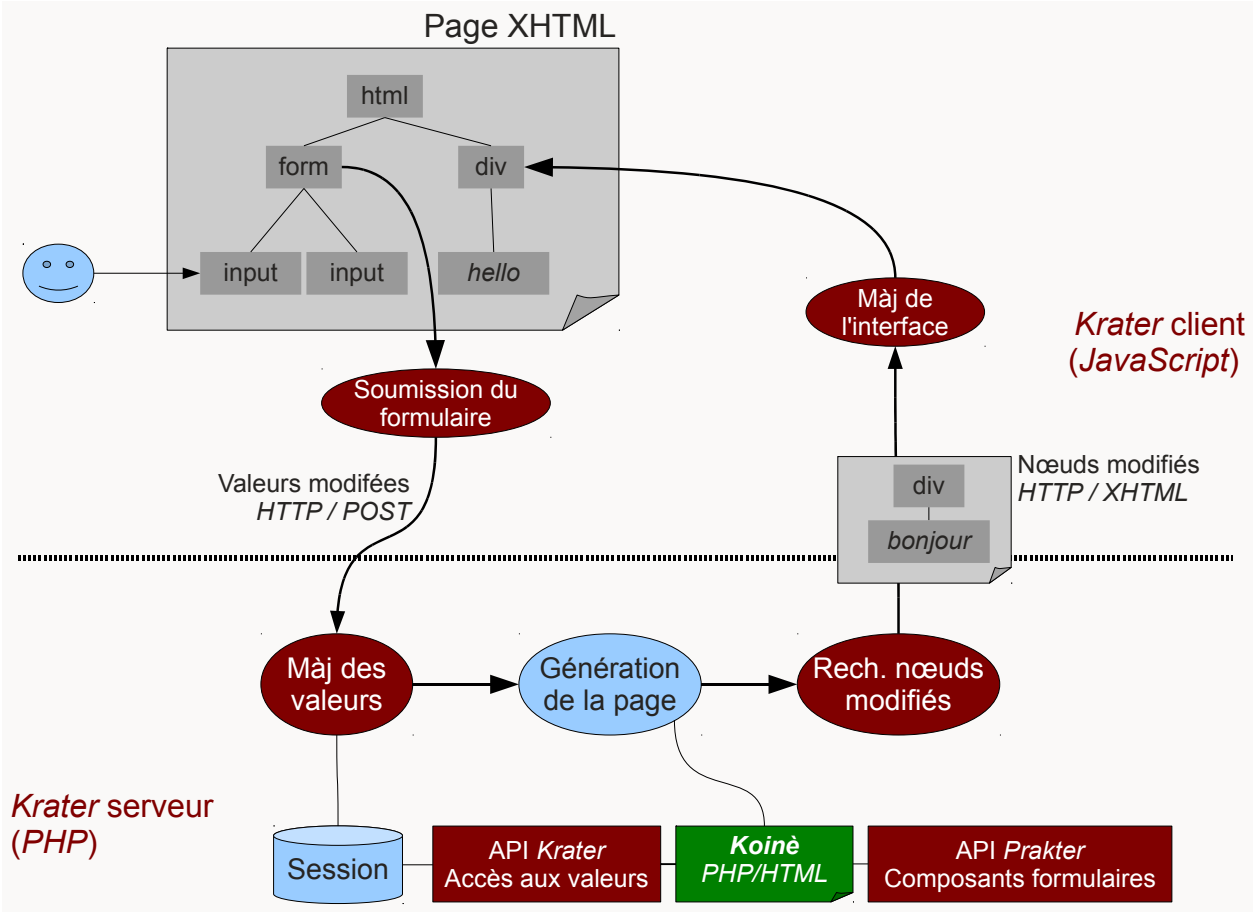


Figure 4.2 – Architecture de Krater-4.

Javascript et d'un champ invisible). On obtient ainsi des variables directement modifiables par l'utilisateur. Ces variables peuvent être relativement complexes ; les tableaux associatifs sont supportés.

Normalement, seules les variables du formulaire qui vient de déclencher le recalcul de la page sont envoyées au serveur. Pour rendre les variables persistantes sur le serveur, nous utilisons le mécanisme des sessions, qui permettent de conserver des variables sur le serveur, automatiquement associées à un client, jusqu'à expiration de la session.

1.4.2 Actions

En outre, chaque intervention de l'utilisateur (interceptée à l'aide d'événements JavaScript) peut être associée à une *action*, c'est à dire un fragment de code PHP qui sera exécuté sur le serveur en réaction à un acte de l'utilisateur.

Par mesure de sécurité, le code PHP de l'action n'est pas visible par le client. Ce dernier ne dispose que d'un identifiant de cette action, que seul le serveur peut

interpréter.

1.4.3 Structures de données évoluées

Cela permet notamment de lier l'interface à des structures de données plus évoluées que les tableaux associatifs, comme des objets, une base de données ou un arbre XML.

Si l'on prend l'exemple d'une base de données, on peut ainsi associer un champ de saisie avec une action « mise à jour du champ correspondant de la base de données ». Un bouton peut être associé avec l'action de créer ou de supprimer une entrée dans une base de données.

Pour un arbre XML, il suffit de lier les actes de l'utilisateur à des fonctions DOM de création, mise à jour et suppression des nœuds de l'arbre XML.

1.4.4 Mise à jour de l'interface

Par défaut, l'envoi d'un formulaire entraîne le rechargement complet de la page. Cela est évidement très désagréable pour l'utilisateur, puisque cela provoque un « clignotement » de la page, et une remise à zéro du défilement et du focus. La technique AJAX[9] permet d'assouplir ce comportement. Au lieu de soumettre normalement les formulaires, ceux-ci sont passés à une fonction Javascript qui se charge d'envoyer les nouvelles données au serveur, de récupérer la nouvelle page XHTML[10], et de mettre à jour les éléments de l'interface affectés, et uniquement eux, par substitution du nouvel élément XHTML à la place de l'ancien. Ce comportement est illustré en figure 4.3. Le focus et la position du curseur sont automatiquement conservés.

Si l'utilisateur modifie l'interface entre l'envoi du formulaire et la réception de la nouvelle page, cette nouvelle page est considérée comme obsolète. La mise à jour est annulée et on attend la nouvelle version de la page. En cas de non réponse du serveur, on relance une requête quelque temps après.

Le serveur peut aussi envoyer à la page XHTML du code Javascript, qui sera automatiquement exécuté à sa réception. Cela peut par exemple permettre de modifier dynamiquement la configuration de la fonction AJAX.

1.4.5 Sécurité

Le vol de session est un risque important pour les utilisateurs d'applications basées sur les sessions. En effet, un pirate peut réussir à récupérer l'identifiant de session d'un utilisateur, et ainsi usurper son identité auprès du serveur. C'est pourquoi l'adresse IP et le nom du navigateur de l'utilisateur sont contrôlés à chaque requête, et doivent rester constants tout au long de la session ; sinon, une nouvelle session est créé.

[9] *Asynchronous Javascript And XML.*

[10] C'est à dire en HTML sous forme de document XML valide.

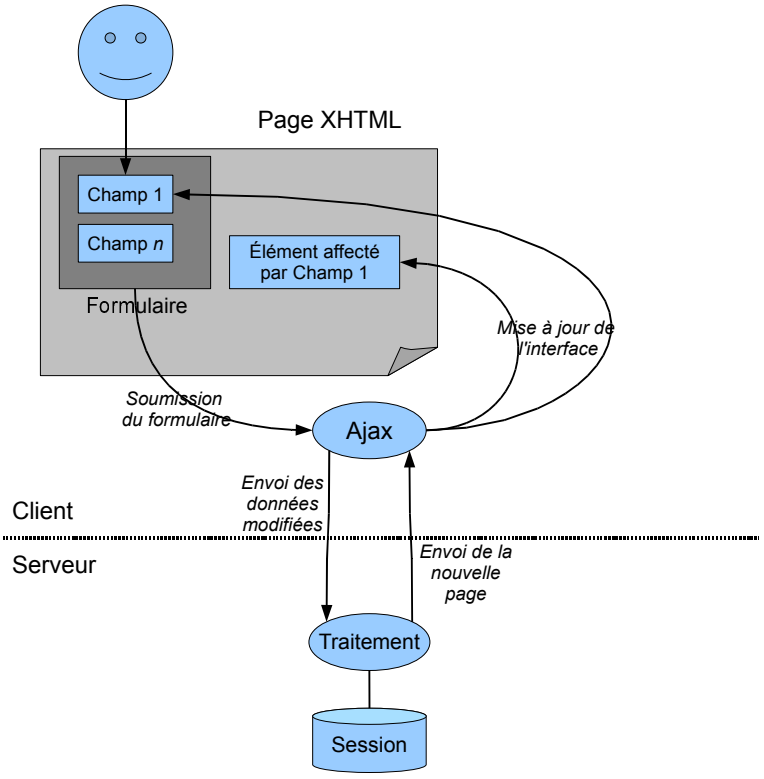


Figure 4.3 – Exemple d'utilisation de Krater-4 lors de la modification d'un champ.

Dans le même souci de sécurité, le contenu des champs de type « mot de passe » est crypté (algorithme MD5) avant l'envoi.

1.4.6 Performances et optimisations

Cette méthode consistant à recalculer la page à chaque modification est particulièrement coûteuse en temps de calcul et en bande passante. C'est pourquoi elle est plus particulièrement destinée à des opérations de prototypage, ou bien en tant qu'interface pour des applications lourdes, avec peu d'utilisateurs, à côté desquelles le temps de calcul consommé par le traitement des pages restera faible. Néanmoins, un certain nombre d'optimisations ont été mises en place.

Tous les champs de données ne causent pas nécessairement des modifications de l'interface. Par exemple, si l'on a trois éléments : un champ de saisie d'identifiant, un champ de saisie de mot de passe, et un bouton de confirmation (lié à un champ caché), seul le dernier élément va provoquer une mise à jour de l'affichage. Certains champs peuvent donc être marqués comme silencieux, les modifications seront notifiées au serveur mais il n'y aura pas de recalcul de la page.

En ce qui concerne la bande passante, une bonne partie de la page XHTML envoyée par le serveur ne sert à rien, étant donné qu'elle n'a subi aucune modification. C'est pourquoi le serveur conserve une copie de toute page envoyée, et n'envoie au client

que le différentiel entre la dernière page stockée (et donc présentement en place sur le client), et la page calculée. Cela permet de ramener les besoins de bande passante de Krater à proximité de ceux d'une application classique basée sur AJAX.

Si l'utilisateur est rapide, ou bien simplement si l'on a configuré un champ de saisie avec un envoi au serveur à chaque caractère entré, le serveur risque d'être inondé de requêtes, d'autant plus inutilement que les réponses du serveur seront considérées comme obsolètes par le client et ignorées. C'est pourquoi il existe un intervalle minimum (paramétrable) entre deux envois au serveur, toute demande d'envoi de données intervenant dans cet intervalle étant mise en attente. À l'issue de cet intervalle, les données en attente sont unifiées (un champ modifié plusieurs fois ne gardera que sa valeur la plus récente) et envoyées au serveur.

À l'inverse, le client doit se manifester régulièrement auprès du serveur pour maintenir la session en vie. C'est pourquoi il existe aussi un intervalle maximum entre deux envois de données au serveur. D'autre part, un intervalle maximum faible peut être utilisé pour synchroniser des clients entre eux, via le serveur, par exemple pour un système de tchat. Ils viendront relever régulièrement des changements, même en dehors de toute action de leur utilisateur.

En ce qui concerne les moteurs de rendu pleinement supportés, c'est actuellement le cas de Gecko (Firefox) et Presto (Opera). Les autres moteurs, à savoir les diverses versions d'Internet Explorer et Webkit (Safari, Chrome) ne sont pour l'instant supportées qu'à travers un mode dégradé, utilisable certes, mais sans les fonctions AJAX de mise à jour de l'interface décrites précédemment.

1.4.7 Boîtes à outils

Par dessus les fonctions de base de Krater, nous avons développé une boîte à outils, Prakter, composée d'éléments graphiques pré-associés à des champs et à un jeu de paramètres par défaut généralement adaptés. Cela permet d'insérer très facilement dans le code des éléments actionnables par l'utilisateur. Par exemple, le code PHP suivant insérera automatiquement un champ de saisie lié à la variable *login* :

```
<?=$k->input('Identifiant', 'login')?>
```

Une seconde boîte à outils permet de gérer des profils d'utilisateur, ainsi que les problématiques de connexion/déconnexion d'un utilisateur, de création de compte avec envoi de courriel de vérification, etc.

Enfin, une troisième boîte à outils permet de partager des informations entre plusieurs sessions, ce qui est particulièrement utile pour un système de tchat.

1.4.8 Réutilisabilité

Cette bibliothèque s'est avérée suffisamment générique pour être utilisée dans d'autres projets, comme SurviTra (décrit en partie 2) et Scientext[11].

Le projet Scientext concerne l'étude du positionnement et du raisonnement dans les écrits scientifiques. Il s'appuie sur un corpus de textes scientifiques annotés morphologiquement par TreeTagger [Sch94] et syntaxiquement par Syntex [BF00] [Bou07]. La consultation du corpus s'appuie sur le moteur ConcQuest [Kra08].

La bibliothèque Krater a été utilisée pour concevoir l'interface d'interrogation du corpus, comprenant la sélection fine de textes (figure 4.4), un éditeur de motifs syn-

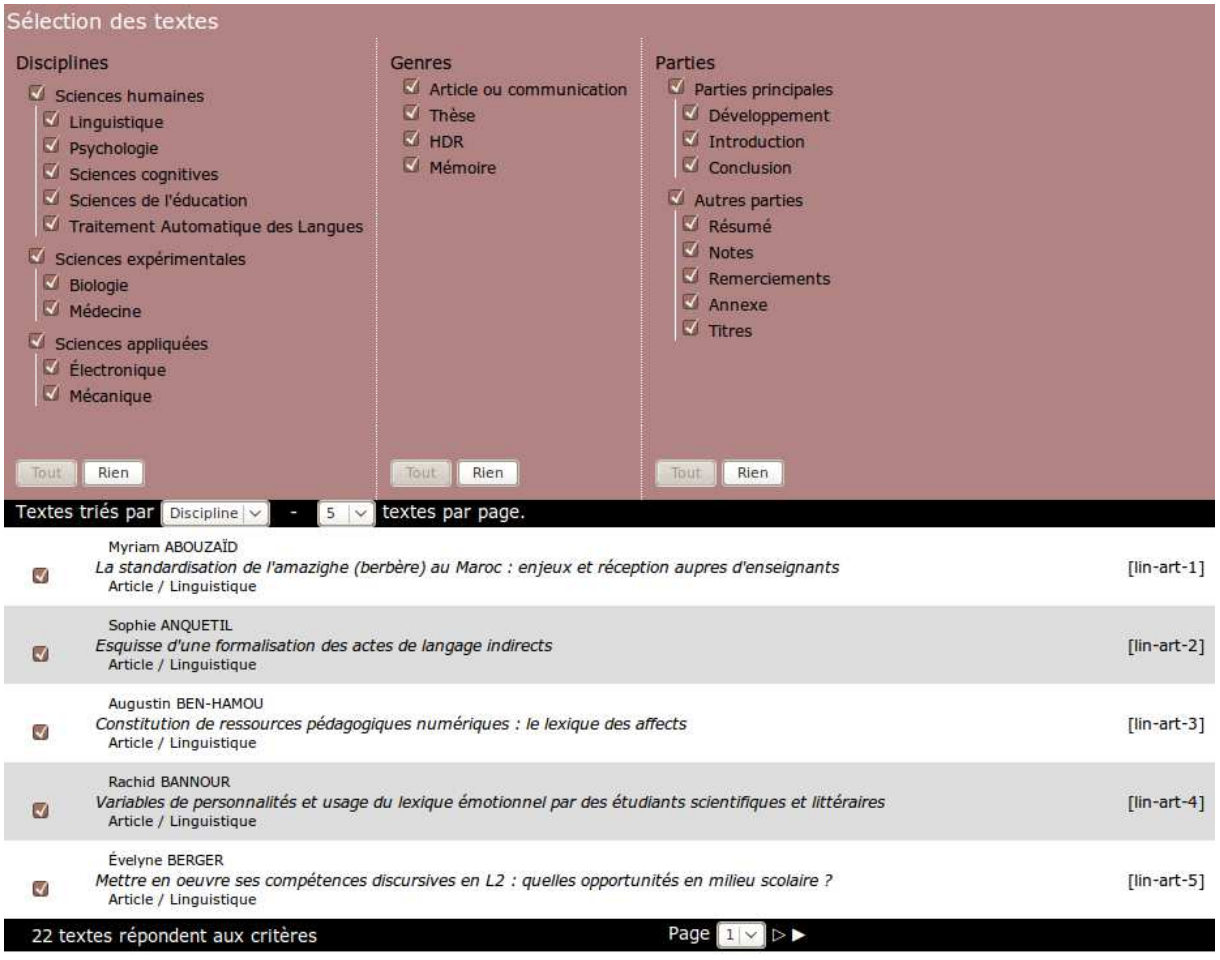


Figure 4.4 – Scientext : choix des textes.

taxiques (selon divers critères, tels que catégorie syntaxique, lemme, forme, expression régulière, et relations syntaxiques entre termes), la présentation des résultats en mode concordancier (figure 4.5), avec visualisation des arbres de dépendance, et enfin l'affichage de statistiques lexicales.

[11] `http://scientext.dynalias.net`

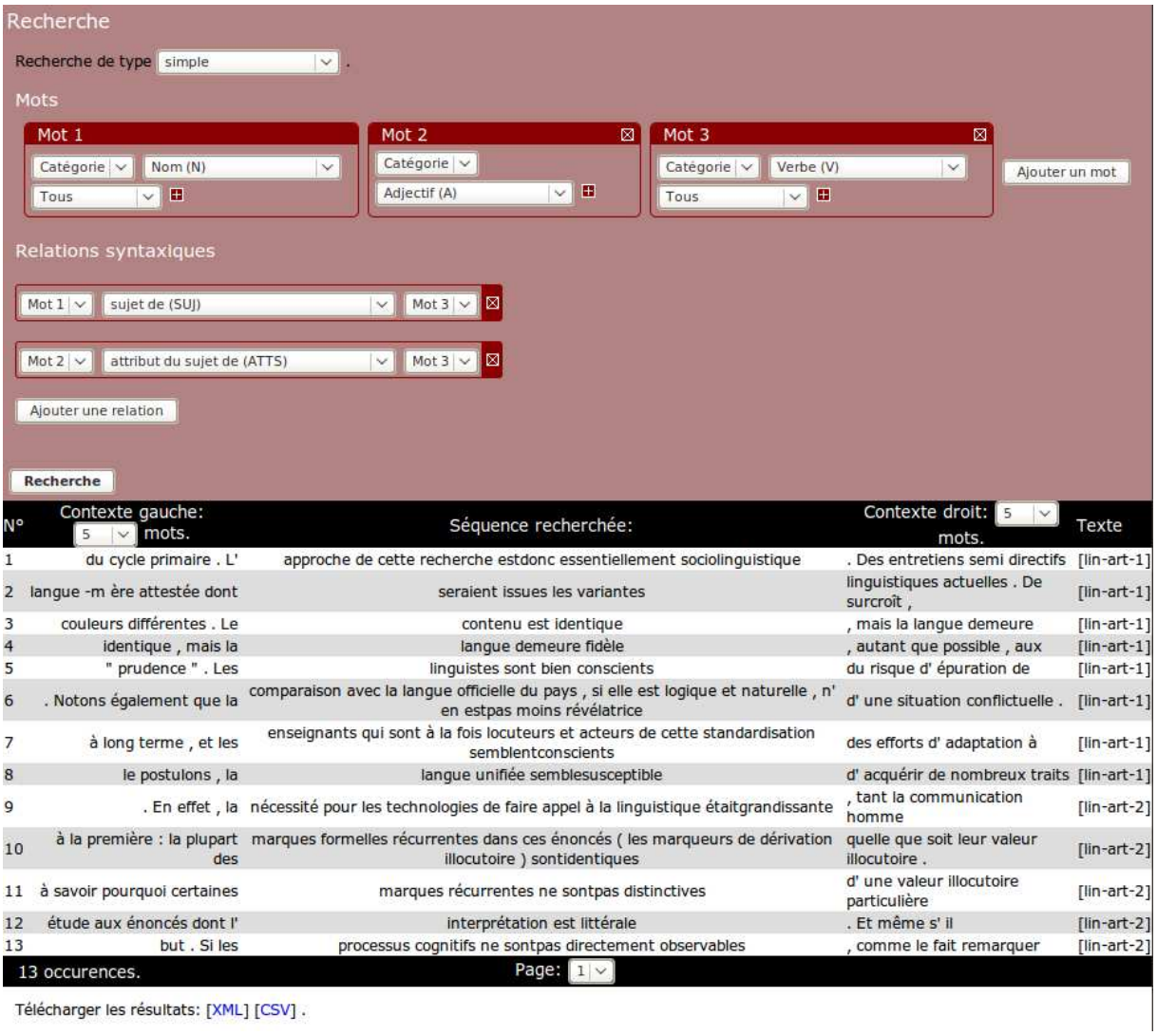


Figure 4.5 – Scientext : recherche syntaxique.

1.4.9 Application multilingue

Le rechargement transparent de la page Web effectué par Krater permet de créer des sites et des applications Web dont la langue peut être modifiée à la volée. Si l'utilisateur modifie la variable indiquant la langue, la version de l'interface prévue pour cette langue s'affichera sans autre modification visible de la page ; le défilement, la sélection, etc. seront conservés. Pour un projet comme SurviTra, par exemple, il était très important que la langue puisse être changée à tout moment sans rien perdre de la recherche de phrases en cours.

2 Koinè : application Web de dialogue en langue seconde

Comme vu en partie 3, Koinè comporte en principe un mode vocal et un mode écrit, complétés par des outils graphiques. Nous présenterons ici les principaux aspects de leur implémentation, en complément de la spécification détaillée donnée en partie 3. Nous présenterons d'abord les ressources auxquelles l'application fait appel (dictionnaires et modèles de langage), puis chacun des modes, et enfin l'outil graphique d'annotation collaborative de documents (pages web, images, plans, etc.) qui les complète.

2.1 Ressources

Dans un premier temps, on vise uniquement la paire anglais-français. Toutefois, les ressources disponibles existent aussi avec une couverture relativement large pour les paires anglais-[allemand, espagnol, italien] et français-[allemand, espagnol, italien], et à terme ces langues seront disponibles.

2.1.1 Dictionnaires

Nous utilisons *sdcv*[12], la version en ligne de commande de StarDict, l'un des principaux logiciels de consultation de dictionnaires, avec deux dictionnaires bilingues bidirectionnels libres, développés par Luc Han et le projet Apertium[13]. Les sorties de ces deux dictionnaires sont factorisées et ramenées à une structure simple (mais la plupart des entrées ne donnent pas plus d'information), associant un mot en langue source et une liste de traductions potentielles en langue cible.

Notre système gagne aussi à pouvoir consulter un dictionnaire de gloses et de synonymes. Nous utilisons la base WordNet, uniquement pour l'anglais.

2.1.2 Modèles de langage

Nos modèles de langage sont construits à partir de la base de données de Wikipedia. Cette base de données est disponible sous forme d'un grand fichier XML pour chaque langue. Ce fichier est tout d'abord divisé en articles, puis chaque article est traité. Si l'article est une discussion, il est « nettoyé » automatiquement de ses méta-informations pour ne conserver que le texte brut. Dans tous les cas, on recherche les informations d'ordre hiérarchique : appartenance à une catégorie, à un portail, etc., et on calcule ainsi pour chaque article la liste de ses « descendants » hiérarchiques. Wikipedia n'est pas strictement hiérarchisée, c'est pourquoi nous contrôlons et complétons manuellement cette ébauche de hiérarchie, en effectuant un classement par thème et sous-thème. Nous obtenons ainsi pour chaque thème et sous-thème, une liste des articles concernés, qui constitue alors notre corpus thématique.

[12] Développé par Evgeniy A. Dushistov et HuZheng : http ://sdcv.sourceforge.net/

[13] Apertium est développé par le groupe de recherche Transducens du Departament de Llenguatges i Sistemes Informàtics de l' Universitat d'Alacant en partenariat avec Prompsit Language Engineering.

Le modèle de langage syntagmatique est construit pour chaque thème et sous-thème, en faisant défiler sur chacun des articles qui le composent une fenêtre de traitement de 100 mots. La distance est alors calculée suivant la formule décrite en partie 3.

Le modèle de langage paradigmatique (n-grammes) est lui aussi construit pour chaque thème et sous thème.

2.2 Mode vocal

Nous avons implémenté une première version de traitement vocal. Cependant, cette implémentation est incomplète, et certains problèmes de qualité et de performances ne sont pas résolus. Le principe est le suivant : l'utilisateur utilise sur son poste local un microphone, relié à un logiciel client, qui transfère le signal sonore vers le serveur, où il est décodé à la volée.

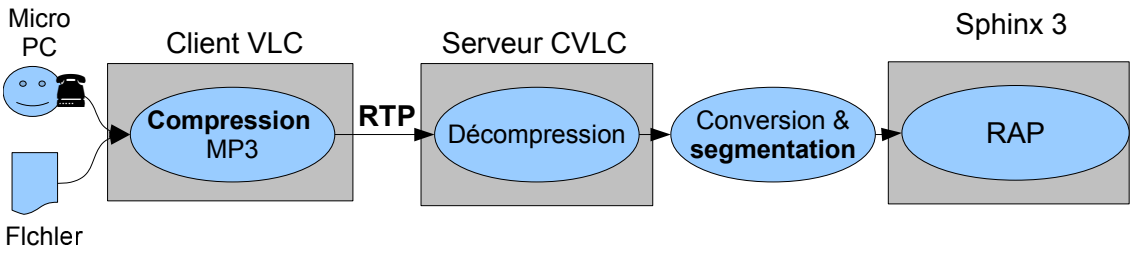


Figure 4.6 – Architecture du système de reconnaissance vocale implémenté.

2.2.1 Sphinx 3

Il s'agit du système universitaire le plus connu, conçu pour être utilisé avec les langages C, Java et Perl (des APIs sont disponibles), qui permet le reconnaissance automatique de parole continue. Il est accompagné d'une boîte à outils permettant de créer des modèles de parole et de langage pour une langue arbitraire. Il a notamment été utilisé dans le cadre d'une thèse sur la reconnaissance de parole pour locuteurs non natifs [TB06].

Nous nous sommes limités à une reconnaissance de l'anglais, pour laquelle il y a le plus de composants disponibles. Mais il existe de nombreuses adaptations pour d'autres langues, de qualité et d'étendue variables, malheureusement souvent propriétaires. Nous utilisons un modèle de langage de 64 000 mots construit à partir de l'*English Gigaword corpus* (textes journalistiques) par Keith Vertanen[14], et le modèle acoustique de parole continue par défaut de Sphinx 3.

[14]http ://www.inference.phy.cam.ac.uk/kv227/lm_giga/

2.2.2 Configuration du client

Tout d'abord, l'utilisateur se connecte à une page Web pour obtenir un numéro de port disponible.

Sur le poste client, le flux audio est compressé au format MP3 et transmis au serveur grâce au logiciel VLC[15], via le protocole RTP[16], sur le port dont le numéro a été donné par le serveur.

En retour, la page Web est régulièrement mise à jour avec le résultat de la reconnaissance vocale.

2.2.3 Conception du serveur

Le serveur utilise une version en ligne de commande de VLC (CVLC) pour récupérer le flux RTP et le convertir (sans perte) en fichier WAV. Nous effectuons une segmentation en tours de parole par détection de silences (200 ms), afin de détecter les fins de tours de paroles. Dès qu'un silence est détecté, le tour de parole est envoyé au moteur de reconnaissance vocale. La reconnaissance étant plus longue que la durée des données à traiter, on peut ainsi avoir plusieurs tours simultanément en cours de reconnaissance. Le résultat de la reconnaissance de chaque tour de parole est affiché dès qu'il est disponible.

2.2.4 Évaluation

Des tests rapides ont été effectués afin d'évaluer les performances de ce système, sans optimisation particulière, mais dans des conditions idéales. En l'occurrence, on s'est servi d'un exposé académique lu par un locuteur natif de manière particulièrement claire, et traitant de questions d'actualité (restant ainsi proche du modèle de langage utilisé). Sur un échantillon de 414 mots, le taux de reconnaissance de mots (WRR) est de 0,6, ce qui est faible, surtout si l'on souhaite utiliser les sorties de ce système comme entrée pour des traitements linguistiques. Toutefois, les mots correctement reconnus sont groupés, formant des îlots de texte corrects ; ainsi on pourrait greffer sur ces groupes les aides présentées en partie 3.

Le processus de reconnaissance vocale est en outre assez lent, puisque sur notre machine serveur[17], la reconnaissance prenait plus de 4 fois de temps que la durée des données à traiter.

Toutefois, dès que l'on remplace l'écrit lu par de la parole spontanée, les performances s'effondrent. En ce qui concerne notre système, les performances deviennent quasiment nulles. Les difficultés liées à la reconnaissance de la parole spontanée ont été présentées par [BMSP92], et sont assez similaires aux problèmes posés par l'écrit

[15] VideoLAN CLient, http ://www.videolan.org/vlc/
[16] Real-Time Transport Protocol, basé sur UDP.
[17] VIA C7 1,5 GHz, voir configuration complète en annexe.

spontané : corrections, pauses, allongements, accentuation, erreurs de prononciation, etc. Des études ont montré que, sur des vocabulaires réduits et pour des tâches spécifiques, il était possible de faire remonter le taux de réussite à 0,56 [SKH+00] (pour du japonais, débat politique, monolocuteur), 0,73 [Yod01] (anglais, multilocuteur, vocabulaire fermé de 523 mots), ou récemment 0,61 [WCS06] (thaï, centre d'appel).

De plus, se pose le problème de la reconnaissance de parole du locuteur non natif. Selon [WS03], un système dont le taux de reconnaissance de mots est de 0,84 passe à 0,52 lorsqu'il est utilisé par des locuteurs non natifs (étude basée sur les données de Verbmobil). Cependant, après adaptation du modèle, le taux remonte à 0,65.

On peut donc considérer que ces deux problèmes sont en voie de résolution, mais que leur réunion reste très problématique, et dépasse largement le cadre de la thèse. De plus, s'agissant d'une thèse CIFRE, l'objectif final reste la réalisation d'un prototype fonctionnel. Enfin, le mode écrit, bien que plus réaliste, posait lui-même des problèmes ; c'est pourquoi nous avons choisi de nous concentrer sur ces problèmes.

2.3 Mode écrit

Devant ces difficultés, nous avons choisi de recentrer l'application sur le mode écrit, ce qui nous permet de rester centré sur notre problématique d'aide au dialogue en langue seconde.

Les modules suivants (spécifiés en partie 3) ont été effectivement implémentés :
- Repérage des mots oraux, à ignorer (émoticons).
- Détection de la langue.
- Détection et standardisation des mots oralisés (abréviations, déformations, fautes involontaires, etc.).
- Détection des mots standards.
- Traduction lexicale (mot à mot), en fonction du contexte.
- Traduction d'un message complet (ou fragment monolingue dans le cas d'un message mutlilingue).

Les résultats de ces modules servent à afficher deux types d'information contextuelle :
- traduction lexicale, contextuelle.
- traduction des messages entiers.

En ce qui concerne les fonctionnalités ne faisant pas intervenir spécifiquement de linguiciel, mais qui répondent à des besoins des utilisateurs que nous avons mis en évidence précédement, trois d'entre elles ont été implémentées :
- sous-titrage lexical.
- correction de messages *a posteriori*.
- sous-dialogue de clarification, attachés à un énoncé ou à une portion, sélectionnée par l'utilisateur, de ce dernier.

2.3.1 Architecture

La conversation en cours est représentée par un arbre XML. Conformément à notre modèle de fonctionnement, l'interface graphique est construite à partir de cet arbre, qui de son côté peut intervenir directement sur les nœuds de l'arbre, entraînant une reconstruction de l'interface, et ainsi de suite.

Lors du processus de construction, un module de base (ordonnanceur) est appelé. Il se charge d'appeler ensuite tous les autres modules en série. Chacun de ces modules reçoit pour unique paramètre l'arbre XML partagé, soit sous forme d'objet DOM (cas des modules internes, c'est à dire développés en PHP et hébergés sur le même serveur), soit sous la forme d'un chemin de fichier (modules intermédiaires, hébergés sur le serveur mais développés en un autre langage), soit enfin sous la forme du fichier XML lui-même (modules externes). Dans ces deux derniers cas, l'appel du modèle se fait par la méthode REST[18], l'une des manières les plus simples d'implémenter un service Web, la communication se faisant par le protocole HTTP, et l'envoi des paramètres (chemin de fichier ou fichier lui-même) par la méthode POST.

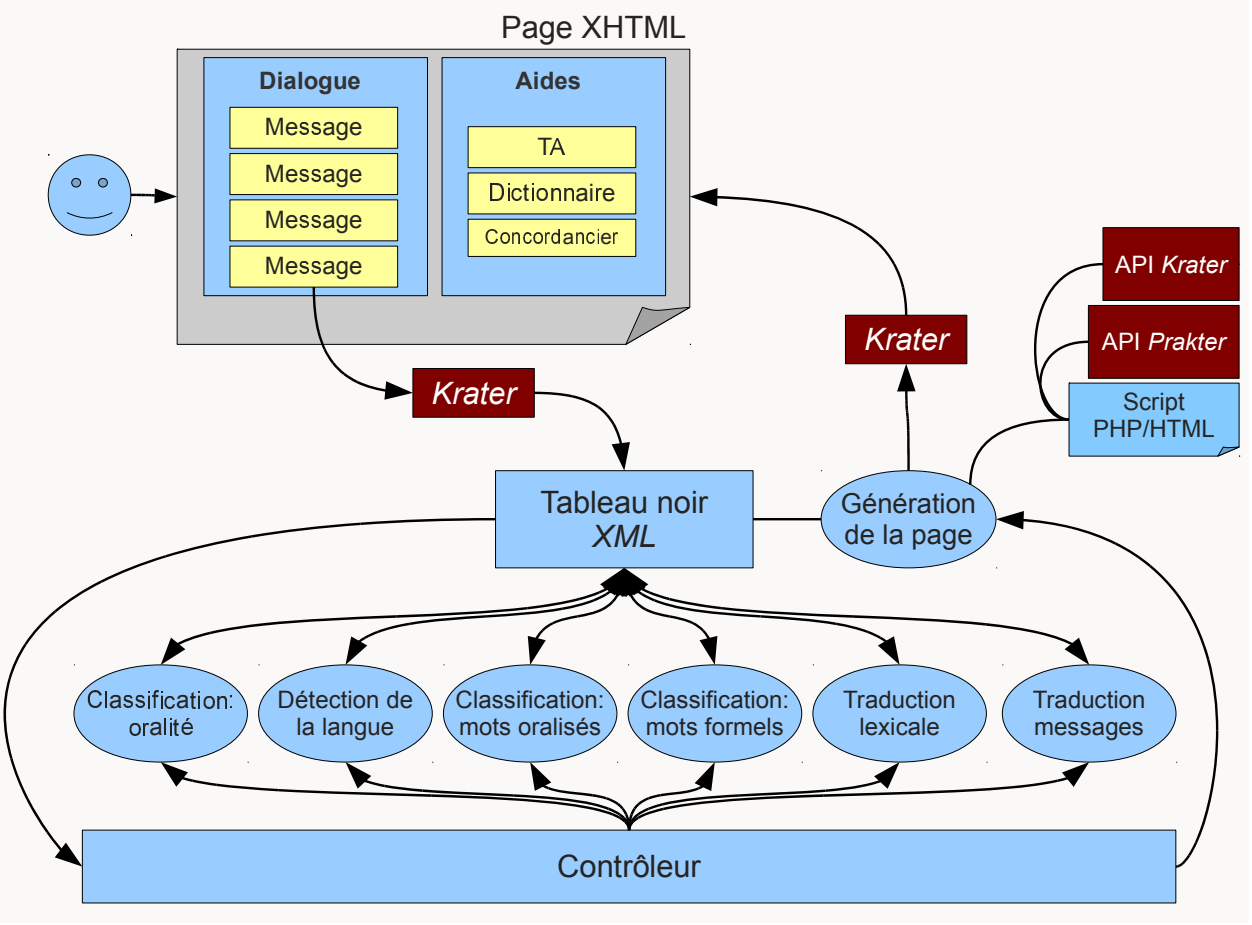


Figure 4.7 – Architecture de Koinè.

[18] Representational State Transfer

2.3.2 Lien chaîne de caractères - objets

D'un côté nous construisons une structure de données, dans laquelle l'utilisateur peut intervenir, à partir du texte qu'il a saisi. Une fois annotée par l'utilisateur, elle ne peut donc plus être déduite du texte, mais doit être conservée. Mais d'autre part, l'utilisateur peut modifier son texte *a posteriori* : on perd alors l'alignement entre la structure et le texte. Par exemple, l'utilisateur peut couper un segment, et le coller ailleurs dans le message ; il s'attend à ce que les annotations effectuées sur ce segment suivent le déplacement. Comment conserver le lien entre la structure annotée par l'utilisateur et le texte modifié ?

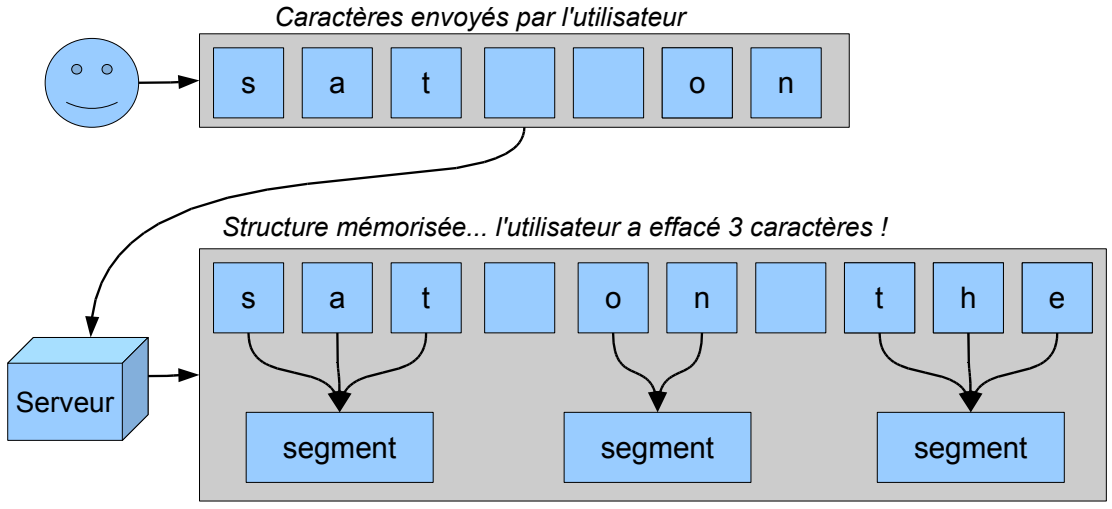


Figure 4.8 – Problème d'alignement.

Nous pourrions créer un champ de saisie avancé, dans lequel les caractères seraient des objets pointant sur la structure, ou, dans le cadre d'un développement avec Krater, toute les actions que l'utilisateur peut effectuer sur le texte seraient liées à des actions visant à modifier la structure en conséquence. Ces actions peuvent être variées : insérer des caractères, en supprimer, en sélectionner, copier, couper, coller, etc. Il y a donc un important travail de développement.

Une autre méthode consiste, pour chaque modification du texte, à réaligner texte et structure. Ainsi, il n'est pas nécessaire de reconstruire un élément « champ de saisie » spécialement pour Koinè.

Nous avons donc fait le choix de cette seconde solution. Nous utilisons l'algorithme de Needleman-Wunsch [NW70], conçu à l'origine pour aligner des séquences d'ADN, mais qui se révèle adapté pour l'alignement de caractères.

2.3.3 Création de conversation et invitation

Koinè étant un outil de « dépannage » linguistique, sa mise en route doit rester la plus simple possible pour les utilisateurs. L'approche de type « messagerie instantanée », avec ses listes de contacts dûment enregistrés, n'est donc pas adaptée. Nous suivons donc une approche assez similaire à celle de l'IRC. N'importe qui peut créer une conversation avec l'identifiant de son choix, sous réserve qu'aucune conversation de ce nom ne soit en cours, simplement en se connectant à une URL de la forme suivante :

```
http:\\koinechat.net?id=identifiant
```

où *identifiant* est un identifiant choisi par l'utilisateur, ou bien généré sur la page d'accueil du projet.

Si la conversation existe déjà, l'utilisateur peut envoyer un message pour demander de la rejoindre. Après accord par au moins un des participants, il peut effectivement rejoindre la conversation.

Ainsi, pour inviter quelqu'un à rejoindre une conversation, il suffit de lui en envoyer l'URL.

2.3.4 Interface

L'interface suit les conventions des outils de tchat (voir figures 4.9 et 4.10). Les messages sont listés de haut en bas, du plus ancien au plus récent. Tout en bas, se trouve le champ de saisie d'un nouveau message. À droite, se trouve le menu contextuel, donnant accès à nos outils.

Certaines informations apparaissent directement dans les messages : problèmes de formalisation et d'orthographe (soulignement rouge), prototypicalité (soulignement noir), signaux (surlignement rouge). Lorsqu'un message est édité (ou créé), il apparaît sous deux formes : d'une part un champ de saisie sans information (aucun soulignement ou surlignement), et d'autre part un champ de visualisation affichant le message avec les informations, tel qu'il apparaitra après envoi (voir figure 4.9). Hormis cela, la partie gauche de l'interface est similaire à celle d'un tchat classique, et Koinè peut donc être utilisé comme tel lorsqu'aucune difficulté ne se présente.

Dans la partie droite de l'interface se trouve le menu contextuel. Son contenu est variable, et dépend du statut du message : nouveau message (figure 4.9), ou déjà envoyé (figure 4.10). Il est organisé en deux parties. En haut (fond noir) se situent tous les outils et informations concernant le message dans son ensemble (historique, langue, thème, sous-titrage, édition). En bas (fond blanc) se situent les informations relatives à la sélection.

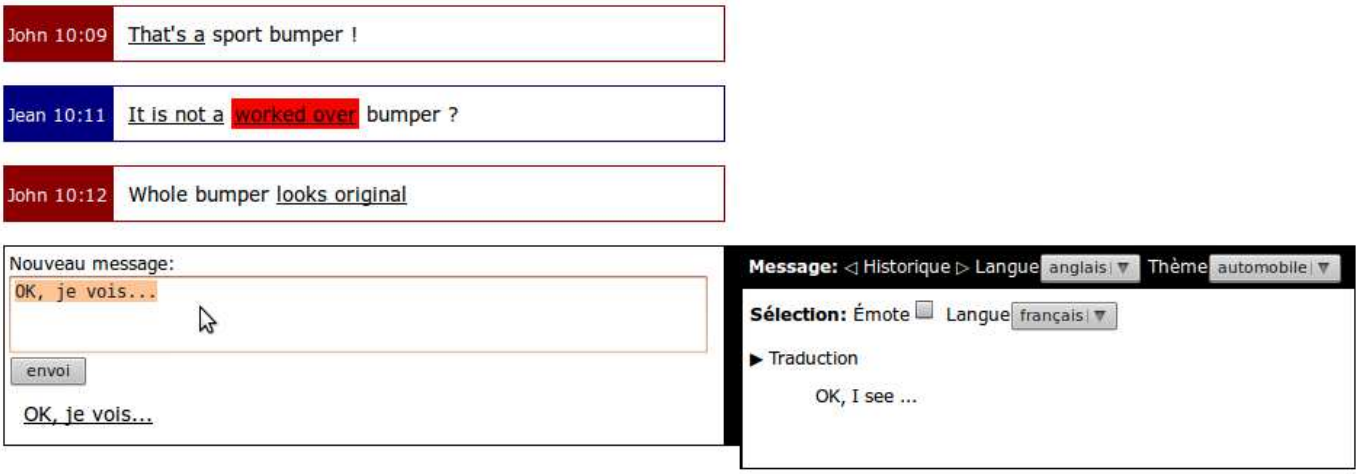


Figure 4.9 – Interface de Koinè en français (non définitive) : composition d'un message.

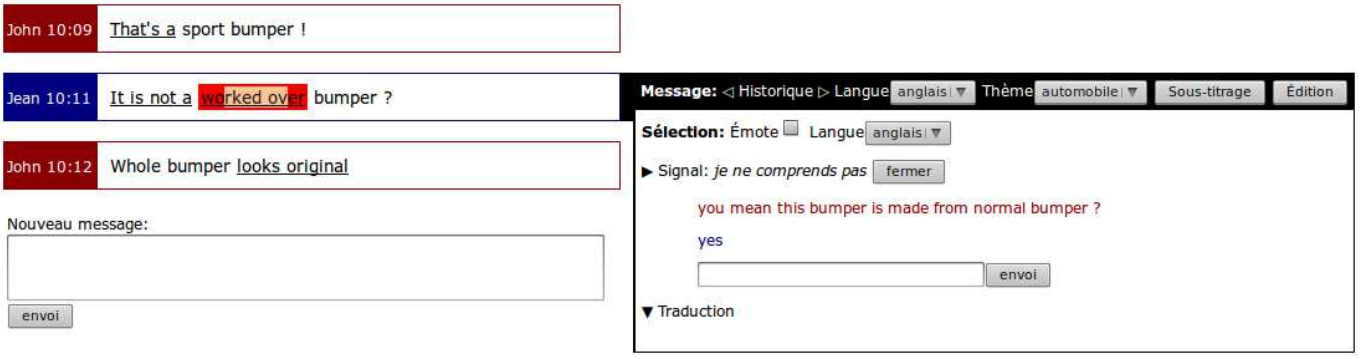


Figure 4.10 – Interface de Koinè en français (non définitive) : sous-dialogue de clarification.

Les informations occupant beaucoup d'espace (par exemple, un signal ou une traduction) sont affichées à la demande. Un symbole de «flèche» pointant vers la droite indique une information masquée, un symbole de «flèche» pointant vers le bas indique une information visible. Un clic sur la flèche permet de passer d'un état à un autre.

Dans la figure 4.10, l'une des informations visible est un sous-dialogue lié à un signal prédéfini «*je ne comprends pas*». Comme nous l'avons vu en partie 3, les signaux permettent de lancer un sous-dialogue de clarification. Dans un tel sous-dialogue, afin d'éviter que l'interface ne devienne trop complexe, les messages peuvent s'enchainer normalement, mais les fonctionnalités sont restreintes : il est juste possible d'envoyer des messages, les informations et les outils ne sont pas disponibles.

2.4 Outils graphiques

Les outils graphiques d'annotation collaborative de documents (pages web, images, plans, etc.) (figure 4.11) s'implémentent tous de la même manière, en superposant trois éléments HTML. Le premier est un cadre contenant une page Web quelconque.

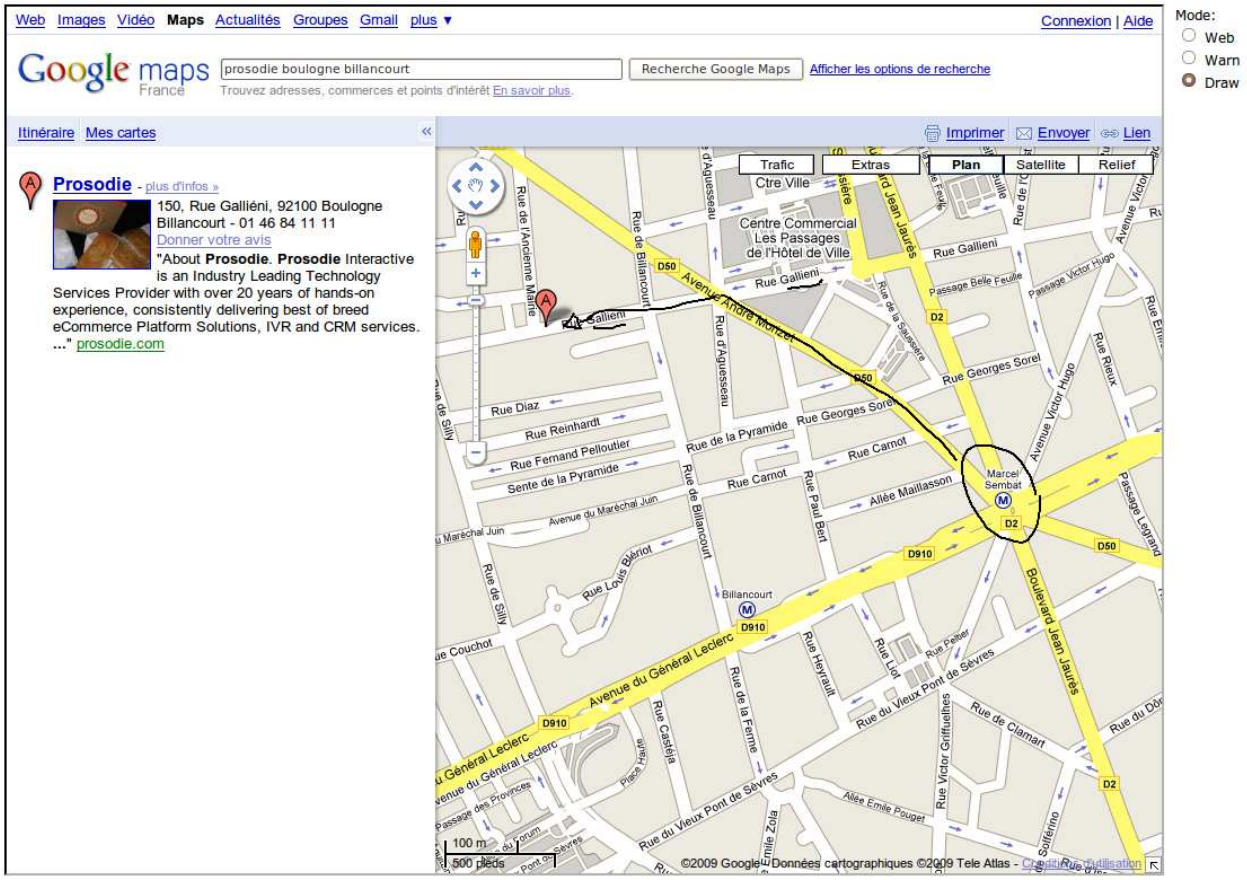


Figure 4.11 – Outil graphique : annotation d'un plan.

Le second, transparent, gère les signaux (structure de donnée partagées entre utilisateurs, comme la conversation). Le troisième, lui aussi transparent, gère les dessins de l'utilisateur (liste de pixels colorés, elle aussi partagée). Un panneau de contrôle permet de passer de l'un à l'autre (à droite sur la figure).

Cela implique que les éléments graphiques en question aient la même taille chez les utilisateurs, indépendamment de leur résolution. On prend donc la résolution du plus petit d'entre eux pour l'étendre aux autres.

Conclusion et perspectives

Apports

Notre travail se situe à la croisée de la linguistique et de l'informatique, et plus particulièrement dans le domaine hybride du traitement automatique de la langue naturelle (TALN). Par conséquent nos apports se trouvent répartis entre ces domaines.

Au niveau linguistique, nous avons contribué à l'étude du dialogue écrit spontané, par la collecte du plus important corpus actuel de tchat (24 millions de mots), et la création d'un outil de navigation et de recherche. Nous avons ainsi pu mener une étude quantitative, qui est venue de compléter les études qualitatives antérieures. Nous avons ainsi pu comparer et mettre en perspective le tchat avec d'autres modalités déjà étudiées par la communauté, et montrer que la différence dans le degré de formalisme que l'on lie généralement à l'opposition écrit / oral ne tient en fait pas à la question de la modalité, mais est une dimension à part entière. L'écrit informel existe, tout aussi bien que la parole formelle, et, sans nier les difficultés inhérentes au traitement de la parole, c'est pour une part importante cet axe formel / informel qui détermine la facilité avec laquelle des traitements automatiques pourront être conçus et mis en œuvre.

Du point de vue du TALN, nos apports sont multiples. Tout d'abord, en ce qui concerne le tchat monolingue, nous avons mis en évidence certaines limites des outils existants, proposé et implémenté des solutions (rectification *a posteriori* des messages, prise en compte du caractère non standard de l'écrit tchaté).

En matière d'aide au dialogue en langue seconde ensuite, qui constitue le cœur de ce travail, nous avons mis en évidence les problèmes, tant du point de vue de l'utilisateur que de celui de la mise en œuvre de solutions. Comme nous l'avons vu, le dialogue spontané, qu'il soit écrit ou parlé, présente des caractéristiques communes, tant au niveau de son caractère improvisé que collaboratif. Cela se traduit par un manque de formalisme, et par des signes qu'une approche fondée sur la norme peut considérer comme des bruits parasites, alors qu'il sont utiles pour évaluer la qualité de la réception et véhiculer l'empathie. On ne peut donc pas s'attendre à ce que les locuteurs acceptent d'adopter une communication purement formelle et dénuée de cette dimension. La présence dans un dialogue d'un locuteur non-natif renforce

certaines de ces caractéristiques. Les tours de parole sont plus décousus, les interventions collaboratives et sous-dialogues de clarification, métalinguistiques ou non, plus fréquents. Les signaux de bonne (ou mauvaise) compréhension sont plus nombreux, de même que les marqueurs d'empathie.

Nous avons ensuite fait des propositions motivées pour l'aide au dialogue en langue seconde, sur différents plans. Sur le plan des outils linguistiques, à l'aide de dictionnaires, de la traduction automatique, d'un concordancier, etc. ; mais aussi à un niveau « méta-utilitaire », en intégrant des possibilités de contrôle et de correction de ces outils ; enfin, nous avons proposé une stratégie visant à faciliter les coopérations métalinguistiques entre les utilisateurs, à travers des sous-dialogues de clarification, ou bien à travers les fonctionnalités de contrôle/correction des aides linguistiques.

Nous avons dans un premier temps été dans le sens d'une vision de la communication en langue étrangère, à médiation forte, négligeant ces aspects pour nous concentrer sur la simple traduction automatique du sens (tchat multilingue, livre de phrases). Il est vrai que cette tendance est générale parmi les outils actuels, qui imposent un formalisme strict aux utilisateurs, les gênant pour véhiculer autre chose que le contenu propositionnel de leurs énoncés. Mais dans un second temps, nous avons tenté d'aborder le dialogue non natif avec une nouvelle approche, basée sur l'observation des stratégies mises en œuvre par les locuteurs en dehors de toute médiation, et consistant à les aider dans la réalisation de ces stratégies en minimisant la médiation : c'est l'approche à médiation faible.

Enfin, d'un point de vue informatique plus général, moins lié au TALN, nous avons développé un prototype, implémentant cette approche sous forme d'application web. L'architecture mise en place dans le cadre du projet Verbmobil s'avère particulièrement générique, et a pu être réutilisée dans ses grandes lignes. Le caractère peu formel des énoncés, et l'aspect fondamentalement collaboratif de leur construction nous ont cependant amené à introduire certaines innovations, notamment la possibilité pour les utilisateurs de modifier leurs énoncés, de corriger les calculs du système aussi bien pour les énoncés dont ils sont auteurs que dans ceux dont ils sont destinataires, et d'associer des sous-dialogues de clarification aux énoncés, éventuellement basés sur les informations métalinguistiques fournies par les composants d'aide.

La réalisation de ce prototype n'a pas été sans difficultés, en particulier du fait de la nécessité d'intégrer une dimension collaborative, nécessairement fortement dynamique, dans une interface. Nous avons toutefois pu ramener ce problème à un cycle simple alternant mise à jour des données par l'interface et mise à jour de l'interface par les données, avec partage de ces données entre les participants. La plate-forme ainsi réalisée s'est avérée suffisamment générique et robuste pour être utilisée dans le cadre d'autres projets (SurviTra, Scientext).

Perspectives

À court terme, nous imaginons trois axes de continuation pour notre travail.

Tout d'abord, certaines fonctionnalités restent à implémenter, comme la vue en concordancier. Certains algorithmes doivent être optimisés : le système met actuellement environ une seconde à réagir aux actions de l'utilisateur, ce qui limite les possibilités d'évaluation de l'outil.

Une fois cela fait, on pourra alors conduire une évaluation du système. Cette évaluation pourrait se faire à deux niveaux : d'une part une évaluation quantitative, en laboratoire, en mesurant le gain de productivité d'un système de tchat « aidé » par rapport à un système sans aides ; et d'autre part, après mise en ligne du système, une évaluation qualitative basée sur l'évaluation des utilisateurs eux-mêmes et l'étude de leur utilisation des outils, afin d'évaluer présicément l'apport de chaque composant à la facilitation du dialogue. En effet, la conception et le développement se sont effectués « en largeur d'abord » ; c'est à dire que nous nous sommes efforcés d'intégrer selon notre principe, le plus simplement possible, des composants couvrant un maximum d'aspects de la communication en langue seconde, sans approfondir aucun d'entre-eux plus que ce qui était requis pour le faire simplement fonctionner. Notre travail ouvre donc logiquement comme première perspective à court terme une étude au cas par cas de chacun de ces composants, afin d'étudier leur impact, leur utilité, et la façon d'en améliorer les résultats. Certains d'entre eux peuvent s'avérer peu utiles, tandis qu'à l'usage certains manques peuvent se faire sentir.

À plus long terme, Koinè pourrait être intégré à une plate-forme de travail collaboratif (*groupware*), préfigurée par les outils de collaboration graphique. Cela se rapprocherait assez du projet *Wave* de Google, qui vise à intégrer une application de messagerie instantanée à d'autres outils en ligne, mais dans le cadre de Koinè, l'accent serait mis sur la dimension multilingue.

Enfin, suivant les évolutions de la reconnaissance automatique de parole, il faudra à l'avenir se pencher à nouveau sur le mode vocal. Dans ce cadre, il serait utile de voir ce que le tchat peut apporter comme modèle de langage pour la reconnaissance vocale, mais aussi pour la construction de modèles de l'information paraverbale et affectionnelle, que les tchatteurs s'échinent à transmettre malgré une modalité qui ne s'y prête guère.

À l'issue de ce travail, nous espérons ainsi avoir entendu l'avertissement que la littérature donne à la science, et contribué à montrer que l'informatique peut rapprocher la machine de l'humain, comme le rêvaient des auteurs tels que René Barjavel, plutôt que le contraire, tant redouté par les auteurs pessimistes comme Ray Bradbury.

Bibliographie

[Ada92] Jean-Michel Adam. *Les textes : types et prototypes. Récit, description, argumentation, explicitation et dialogue.* Fernand Nathan, 1992.

[AJ98] M. Akbar and Caelen J. Parole et traduction automatique : le module de reconnaissance RAPHAEL. *Proceedings of the 17th international conference on Computational linguistics*, 1998.

[BB94] C. Boitet and H. Blanchon. Multilingual Dialogue-Based MT for Monolingual Authors : the LIDIA Project and a First Mockup. *Machine Translation*, 9 :99–132, 1994.

[BB00] Claire Blanche-Benveniste. *Approches de la langue parlée en français.* Ophrys, 2000.

[BB04a] Hervé Blanchon and Christian Boitet. Deux premières étapes vers les documents auto-explicatifs. In *Proc TALN 2004*, pages 47–50, Fès, Maroc, avril 2004.

[BB04b] Hervé Blanchon and Christian Boitet. Les documents auto-explicatifs : une voie pour offrir l'accès au sens aux lecteurs. In *CIDE 7 : Approches Sémantique du Document Électronique*, La Rochelle, jun 2004.

[BBB+07] Christian Boitet, Pushpak Bhattacharyya, Etienne Blanc, Sanjay Meena, Sangharsh Boudhh, Georges Fafiotte, Achille Falaise, and Vishal Vacchani. Building Hindi-French-English-UNL resources for SurviTra-CIFLI, a linguistic survival system under construction. In *SNLP 2007*, Pattaya, Thaïlande, 2007.

[BBBD+79] C. Blanche-Benveniste, B. Borel, J. Deulofeu, J. Durand, A. Giacomi, C. Loufrani, B. Meziane, and N. Pazery. Des grilles pour le français parlé. *Recherches sur le Français Parlé*, pages 163–206, 1979.

[BBC99] H. Blanchon, C. Boitet, and J. Caelen. Participation Francophone au Consortium C-STAR II. *La tribune des industries de la langue et du multimédia / Linguistic engineering and multimedia tribune*, 31–32 :15–23, août-décembre 1999.

[BBF+01] L. Besacier, H. Blanchon, Y. Fouquet, J.P. Guilbaud, S. Helme, S. Mazenot, D. Moraru, and D. Vaufreydaz. Speech translation for french in the Nespole ! european project. In *Eurospeech'01*, pages 1291–1294, Aalborg, Danemark, septembre 2001.

[BBL04] H. Blanchon, C. Boitet, and Besacier L. Spoken dialogue translation system evaluation : Results, new trends, problems and proposals. In

Proc. IWSLT 2004 (ICLSP 2004 Satellite Workshop), pages 95–102, Kyoto, Japon, septembre-octobre 2004.

[BCD+04] Christophe Benzitoun, Estelle Campione, José Deulofeu, Sandrine Henry, Frédéric Sabio, Sandra Teston, André Valli, and Jean Véronis. L'analyse syntaxique de l'oral : problèmes et méthodes. *Journée d'Etude de l'ATALA sur l'annotation syntaxique de corpus*, mai 2004.

[BDM+00] M. Bagein, T. Dutoit, F. Malfrere, V. Pagel, A. Ruelle, N. Tounsi, and D. Wynsberghe. The euler project : an open, multi-lingual and multi-platform text-to-speech system. In *Proceedings of ProRISC'2000*, pages 193–197, Veldhoven (Pays-Bas), dec 2000.

[Bea27] Charles Beaulieux. *Histoire de l'orthographe française, tome I : Formation de l'orthographe des origines au milieu du XVIe siècle.* Champion, Paris, 1927.

[Bei01] Michael Beißwenger, editor. *Chat-Kommunikation. Sprache, Interaktion und Sozialität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart, Allemagne, 2001.

[Ber04] Vincent Berment. *Méthodes pour informatiser des langues et des groupes de langues « peu dotées »*. PhD thesis, Université Joseph Fourier – Grenoble 1, juin 2004.

[BF00] D. Bourigault and C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25 :131–151, 2000.

[BFS+07] P. Bouillon, G. Flores, M. Starlander, N. Chatzichrisafis, M. Santaholma, N. Tsourakis, M. Rayner, and B.A. Hockey. A bidirectional grammar-based medical speech translator. In *Proceedings of the ACL Workshop on Grammar-based Approaches to Spoken Language Processing*, pages 41–48, Prague, République tchèque, 2007.

[BG00] C. Boitet and J.-P Guilbaud. Analysis into a formal task-oriented pivot without clear abstract semantics is best handled as "usual" translation. In *Proc. ICSLP 2000*, pages 436–439, Pékin, Chine, 2000.

[BHN+00] Anton Batline, Richard Huber, Heinrich Niemann, Elmar Nöth, Jörg Spilker, and Kerstin Fischer. *Verbmobil : Foundations of Speech-to-Speech Translation*, chapter The Recognition of Emotion, pages 635–658. Springer-Verlag, 2000.

[BL02] Didier Bourigault and Guiraude Lame. Analyse distributionnelle et structuration de terminologie, application à la construction d'une ontologie documentaire du droit. *TAL*, 43, 2002.

[Bla94] H. Blanchon. *LIDIA-1 : une première maquette vers la TA interactive « pour tous »*. PhD thesis, Université Joseph Fourier - Grenoble 1, janvier 1994.

[Bla04a] Hervé Blanchon. Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de tao de l'écrit et de l'oral (HDR). Technical report, Université Joseph Fourier, Grenoble, 2004.

[Bla04b] Hervé Blanchon. Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de tao de l'écrit et de l'oral (HDR), 2004.

[Bla05] Étienne Blanc. About the french enconverter and the french deconverter. In *UNL : Advances in Theory and Applications, Research on Computing Science*, pages 157–166, Instituto Politécnico Nacional, Mexico, Mexique, 2005.

[BLC+05] L. Besacier, V.-B. Le, E. Castelli, S. Sethsery, and L. Protin. Reconnaissance automatique de la parole pour les langues peu dotées : Application au vietnamien et au khmer. In *TALN 2005*, Dourdan, juin 2005.

[BMSP92] John Butzberger, Hy Murveit, Elizabeth Shriberg, and Patti Price. Spontaneous speech effects in large vocabulary speech recognition applications. In *HLT '91 : Proceedings of the workshop on Speech and Natural Language*, pages 339–343, Morristown, New-Jersey, États-Unis, 1992. Association for Computational Linguistics.

[Boi98] Christian Boitet. Problèmes scientifiques intéressants en traduction de parole. In *International Conference on Natural Language Processing and Industrial Applications*, Moncton, Nouveau-Brunswick, Canada, 1998.

[Bou93] Daniel Bougnoux. *Sciences de l'information et de la communication.* Larousse, collection Textes essentiels, 1993.

[Bou07] Didier Bourigault. Un analyseur syntaxique opérationnel : Syntex (HDR). Technical report, Université Toulouse le Mirail, 2007.

[BRC+05] P. Bouillon, M. Rayner, N. Chatzichrisafis, B.A. Hockey, M. Santaholma, M. Starlander, Y. Nakao, K. Kanzaki, , and Isahara H. A generic multi-lingual open source platform for limited-domain medical speech. In *10th Conference of the European Association for Machine Translation (EAMT)*, Budapest, Hongrie, 2005.

[Bre06] Joan Bresnan. Is syntactic knowledge probabilistic ? experiments with the english dative alternation. In *International Conference on Linguistic Evidence*, Tuebingen, Allemagne, feb 2006.

[BS07] Michael Beißwenger and Angelika Storrer. *Corpus Linguistics. An International Handbook.*, chapter Corpora of Computer-Mediated Communication. Mouton de Gruyter, 2007.

[BT95] Christian Boitet and Mutsuko Tomokio. Ambiguities and ambiguity labelling : towards ambiguity databases. In *RANLP'95 (Recent Advances in NLP)*, volume 1/1, pages 13–26, Tzigov Chark, Bulgarie, septembre 1995.

[Cae02] Jean Caelen. Modèles formels de dialogue. In *Actes des 2èmes assises du GdR I3, Information, Interaction Intelligence*, pages 31–58. Cépadues, 2002.

[Cah99] Gérald Cahen. Préface. *Autrement*, 182, janvier 1999.

[CH04] Tom Cobb and Marlise Horst. *Vocabulary in Second Language*, chapter Is There Room for an Academic Word List in French ?, pages 15–38. Benjamins, Amsterdam, Pays-Bas, 2004.

[CL98] David Camballo and Joseph Lo. The IRC prelude. `http://www.irchelp.org/irchelp/new2irc.html`, 1998.

[CN06] Jean Caelen and Hoá Nguyen. Traitement des incompréhensions et des malentendus en dialogue homme-machine. In *Actes de TALN 06*, Louvain, Belgique, avril 2006.

[CPBB08] Huynh Cong-Phap, Christian Boitet, and Hervé Blanchon. SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora. In *Actes de LREC-08*, page 6, Marrakech, Maroc, mai 2008.

[Cry01] David Crystal. *Language and the Internet*. Cambridge University Press, Cambridge, Royaume-Uni, 2001.

[Del06] Nicole Delbecque, editor. *Linguistique cognitive, comprendre comment fonctionne le langage*. de Boeck, 2006.

[DL93] T. Dutoit and H. Leich. *Speech Communication*, chapter An Introduction to Text-To-Speech synthesis based on an MBE Re-Synthesis of the Segments Database. Elsevier Publisher, 1993.

[DP96] T. Dutoit and V. Pagel. Le projet MBROLA : Vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. In *Actes des Journées d'Étude sur la parole*, pages 441–444, Avignon, 1996.

[DR01] Luigi Canali De Rossi. Free demo multilingual chat systems. `http://www.masternewmedia.org/2001/12/31/free\_demo\_multilingual\_chat\_systems.htm`, 2001.

[DS95] Oswald Ducrot and Jean-Marie Schaeffer, editors. *Nouveau dictionnaire encyclopédique des sciences du langage*. Points, 1995.

[Dut96] T. Dutoit. *An Introduction to Text-To-Speech Synthesis*. Kluwer Academic Publishers, 1996.

[Eul01] Eulogos. Corpus di conversazioni da chat-line in lingua italiana, da registrazioni effettuate nel primo trimestre 1998. `http://www.intratext.com/X/ITA0192.HTM`, 2001.

[Faf04a] Georges Fafiotte. Building and sharing multilingual speech resources, using erim generic platforms. In *Proc. COLING-MLR 2004*, 2004.

[Faf04b] Georges Fafiotte. Interprétariat à distance et collecte de dialogues spontanés bilingues, sur une plate-forme générique multifonctionnelle. In *Actes TALN 2004*, Fès, Maroc, avril 2004.

[Fal04] Achille Falaise. Premier pas vers une TA interactive pour le tchat, mémoire de master de recherche. Master's thesis, Université Joseph Fourier, Grenoble, 2004.

[Fal05] Achille Falaise. Constitution d'un corpus de français tchaté. In *Actes de RÉCITAL 2005*, Dourdan, juin 2005.

[FBSZ04] G. Fafiotte, C. Boitet, M. Seligman, and C.-Q. Zong. Collecting and sharing spontaneous speech corpora : the chinfadial experiment. In *Proc. LREC 2004*, Lisbonne, Portugal, mai 2004.

[FFG09] Georges Fafiotte, Achille Falaise, and Jérôme Goulian. CIFLI-SurviTra, deux facettes : démonstrateur de composants de TA fondée sur UNL, et phrasebook multilingue. In *Actes de TALN 2009*, Senlis, juin 2009.

[Fir96] Alan Firth. The discursive accomplishment of normality : on 'lingua franca' english and conversation analysis. *Journal of Pragmatics*, 26 :237–259, 1996.

[FKP06] C. Fairon, J. Klein, and S. Paumier. *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation.* Presses universitaires de Louvain, Louvain, Belgique, 2006.

[FM07] Eric N. Forsyth and Craig H. Martell. Lexical and discourse analysis of online chat dialog. *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 19–26, septembre 2007.

[FOG] Portail du gouvernement, mode d'emploi des forums. `http://www.forums.gouv.fr/infos.php3?id\_article=6`.

[Fuc00] Catherine Fuchs. *Les ambiguïtés du français.* Ophrys, 2000.

[GAD02] V. Gedner and M. Adda-Decker. Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques. *Actes des XIVe Journées d'Etude sur la Parole*, 2002.

[GD07] R. Kelly Garrett and James N. Danziger. IM=Interruption Management ? Instant Messaging and Disruption in the Workplace. *Journal of Computer-Mediated Communication*, 2007.

[GdNV04] E. Guimier de Neef and J. Véronis. 1 pw1 sr la kestion ;-). In *Journée d'étude de l'ATALA, Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, Paris, 2004.

[GL05] Marta Gonzalez-Lloret. *The Consequences of Mobility : Linguistic and sociocultural contact zones*, chapter Reconstructing NS/NNS communication. Roskilde University, Danemark, 2005.

[Gnu00] C. Gnutzmann. *Routledge encyclopedia of language teaching and learning*, chapter Lingua franca, pages 356–359. Routledge, Londres, Royaume-Uni, 2000.

[Goo81] Ch. Goodwin. *Conversational Organization : Interaction between Speakers and Hearers.* Academic Press, New-York, États-Unis, 1981.

[Gus99] Ela Gusakowska. Business english and cross cultural problems. *Humanising Language Teaching*, 8, décembre 1999.

[HBS08] Cornelia Hülmbauer, Heike Böhringer, and Barbara Seidlhofer. Introducing english as a lingua franca (ELF) : Precursor and partner in intercultural communication. *Synergies Europe*, 3 :25–36, 2008.

[Her99] S. Herring. Interactional coherence in CMC. *Journal of computer-Mediated Communication*, 4, 1999.

[icq] The ICQ Story. `http://www.icq.com/info/icqstory.html`.

[IUI06] Emi IZUMI, Kiyotaka UCHIMOTO, and Hitoshi ISAHARA. Measuring intelligibility of japanese learner english. In *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006*, pages 476–487, Turku, Finlande, 2006. Springer.

[Jak63] Roman Jakobson. *Essai de linguistique générale (tome I)*. Éditions de Minuit, Paris, 1963.

[Jak73] Roman Jakobson. *Essai de linguistique générale (tome II)*. Éditions de Minuit, Paris, 1973.

[JO0] Journal Officiel du 5 avril 2006 - Déclaration du terme «dialogue en ligne» par la Commission Générale de Terminologie et de Néologie.

[JO9] Journal Officiel du 16 mars 1999 - Déclaration du terme «causette» par la Commission Générale de Terminologie et de Néologie.

[JvH00] Susanne J. Jekat and Walther v. Hahn. *Verbmobil : Foundations of Speech-to-Speech Translation*, chapter Multilingual Verbmobil-Dialogs : Experiments, Data Collection and Data Analysis. Springer-Verlag, 2000.

[Kac92] B. B. Kachru. *The other tongue (2nd edition)*, chapter Models for non-native Englishes, pages 48–74. University of Illinois Press, Urbana and Chicago, États-Unis, 1992.

[Kel87] Jeff Kell. Relay : Past, present, and future. In *Actes de NETCON*, Nouvelle-Orléans, États-Unis, 1987.

[KNK00] Andreas Klüter, Alassane Ndiaye, and Heinz Kirchmann. *Verbmobil : Foundations of Speech-to-Speech Translation*, chapter Verbmobil From a Software Engineering Point of View : System Design and Software Integration, pages 635–658. Springer-Verlag, 2000.

[KO05] Catherine Kerbrat-Orecchioni. *Le discours en interaction*. Armand Colin, 2005.

[Kra97] Susanne Krause. Kommunikation im Internet. Chatten im IRC als Form des Gesprächs. Master's thesis, Universitäts-Gesamthochschule Siegen, Allemagne, 1997.

[Kra08] Olivier Kraif. Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest. In *9ème Journées d'analyse statistique des données textuelles, JADT 2008*, volume 2, pages 625–634. Presses universitaires de Lyon, 2008.

[KW97] E. Keller and S. Werner. Automatic intonation extraction and generation for french. In *Proc. 14th CALICO Annual Symposium*, West-Point, États-Unis, 1997.

[KZ98] E. Keller and B. Zellner. Motivations for the prosodic predictive chain. In *Proc. ESCA Symposium on Speech Synthesis*, pages 137–141, Jenolan Caves, Australie, 1998.

[LaP08] Jeff LaPorte. Global instant messaging market share - open data. http://billionsconnected.com/blog/2008/08/global-im-market-share-im-usage/, 2008.

[Lec] Jacques Leclerc. L'aménagement linguistique dans le monde. http://www.tlfq.ulaval.ca/axl/.

[lgu06] Europeans and their languages. *Special Eurobarometer*, 2006.

[LT00] Guillaume Latzko-Toth. L'Internet Relay Chat : un cas exemplaire de dispositif sociotechnique. *COMMposite*, janvier 2000.

[LT01a] Guillaume Latzko-Toth. L'Internet Relay Chat : un dispositif sociotechnique riche d'enseignements. In *Actes du XIIe Congrès national des sciences de l'information et de la communication*. UNESCO, 2001.

[LT01b] Guillaume Latzko-Toth. Un dispositif construit par ses utilisateurs ? le rôle structurant des pratiques de communication dans l'évolution technique de l'Internet Relay Chat. In *Actes du IIIème colloque international sur les usages et services des télécommunications*, 2001.

[LWL97] A. Lavie, A. Waibel, and L. Levin. Janus III : speech-to-speech translation in multiple languages. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1997.

[Lê06] Viêt Bac Lê. *Reconnaissance automatique de la parole pour des langues peu dotées*. PhD thesis, université Joseph Fourier, 2006.

[Mag] M. Maghsoodnia. Instant messaging, ready for prime time ? http://www.wsta.org/publications/articles/1202_article04.html.

[Mal08] Artur Malek. Czerwcowe wyniki megapanelu - najpopularniejsze witryny w polsce. *Internet Standard*, 2008.

[MBC+03] N. Mana, S. Burger, R. Cattoni, L. Besacier, V. Maclaren, J. Mc Donough, and F. Metze. The Nespole ! VoIP Corpora in Tourism and Medical Domains. In *EUROSPEECH 2003*, Genève, Suisse, septembre 2003.

[Mei96] Christiane Meierkord. *Englisch als Medium der interkulturellen Kommunikation, Untersuchungen zum non-native/non-native-speaker-Diskurs*. Peter Lang, Francfort sur le Main, Allemagne, 1996.

[Mei00] Christiane Meierkord. *Gesprächsforschung : neue Entwicklungen*, chapter Interpreting successful lingua franca interaction. An analysis of non-native-/non-native small talk conversations in English. 2000.

[Mil99] Jeremie Miller. Open real time messaging system. http://slashdot.org/articles/99/01/04/1621211.shtml, janvier 1999.

[Min07] Andrew Min. The history of the instant messengers, from IRC to Pidgin. *Free Software Magazine*, juin 2007.

[MM] Florence Myles and Rosamond Mitchell. FLLOC projects. http://www.flloc.soton.ac.uk.

[MMS+02] F. Metze, J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari,

N. Mana, and E. Pianta. The Nespole ! speech-to-speech translation system. In *Proceedings of HLT-2002 Human Language Technology Conference*, San Diego, Etats-Unis, mars 2002.

[Mon05] L. Mondada. *Second language conversations*, chapter Ways of "doing being plurilingual" in international work meetings, pages 18–39. Continuum, Londres, Royaume-Uni, 2005.

[MP00] Edmond Marc and Dominique Picard. *Relations et communications interpersonnelles*. Dunod, Paris, 2000.

[NW70] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48 :443–453, 1970.

[OC91] S. Oviatt and P. Cohen. Discourse structure and performance efficiency in interactive and non interactive spoken modalities. *Computer, Speech and Language*, pages 297–326, 1991.

[Oik93] Jarkko Oikarinen. IRC history... `http://www.the-project.org/history.html`, 1993.

[OQL] Grand dictionnaire terminologique de l'Office Québecois de la Langue Française. `http://www.granddictionnaire.com/btml/fra`.

[Owe81] M. Owen. Conversational units and the use of 'well'. *Conversation in Discource*, pages 99–116, 1981.

[Pas99] Madeleine Pastinelli. Ethnographie d'une délocalisation virtuelle : le rapport à l'espace des internautes dans les canaux de "chat". *Terminal*, 79 :41–60, 1999.

[Pie03a] Isabelle Pierozak. *Le français tchaté. Une étude en trois dimensions (sociolinguistique, syntaxique et graphique) d'usages IRC*. PhD thesis, Université d'Aix-Marseille I, 2003.

[Pie03b] Isabelle Pierozak. Le « français tchaté » : un objet à géométrie variable ? *Langage et Société*, 104 :123–144, 2003.

[Pit05] Marie-Luise Pitzl. Non-understanding in english as a lingua franca : examples from a business context. *Views*, 14/2 :50–71, 2005.

[Pol93] Szatrowski Polly. A structural analysis of japanese discourse. an investigation into the strategy of invitations. *Frontier series*, 5, 1993.

[Rau02] E. Rauscher. Des glyphes de pierre qui refont sens. *Science et Vie*, HS 219 :136–147, 2002.

[RBB+08] Manny Rayner, Pierrette Bouillon, Jane Brotanek, Glenn Flores, Sonia Halimi, Beth Ann Hockey, Hitoshi Isahara, Kyoko Kanzaki, Elisabeth Kron, Yukie Nakao, Marianne Santaholma, Marianne Starlander, and Nikos Tsourakis. The 2008 MedSLT system. In *proceedings of Coling 2008 Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, Manchester, Royaume-Uni, août 2008.

[RBC+06] Manny Rayner, Pierrette Bouillon, Nikos Chatzichrisafis, Marianne Santaholma, Marianne Starlander, Beth Ann Hockey, Yukie Nakao, Hitoshi Isahara, and Kyoko Kanzaki. MedSLT : A limited-domain unidirectional grammar-based medical speech translator. In *Workshop on*

Medical Speech Translation, at North American Chapter of the Association for Computational Linguistics - Human Language Technologies '06, 2006.

[Rog51] Carl R. Rogers. *Client-centered therapy. It current practice, implications, and theory.* Houghton Mifflin Company, 1951.

[Ros73] Eleanor Rosch. Natural categories. *Cognitive Psychology*, pages 328–350, 1973.

[Ros75] Eleanor Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology : General*, pages 192–233, septembre 1975.

[San90] Beatrice Santorini. Part-of-speech tagging guidelines for the penn treebank project. Technical report, University of Pennsylvania, 1990.

[Sau16] Ferdinand de Saussure. *Cours de linguistique générale*. Payot, 1916.

[SBC+05] M. Starlander, P. Bouillon, N. Chatzichrisafis, M. Santaholma, M. Rayner, B.A. Hockey, H. Isahara, K. Kanzaki, and Y. Nakao. Practicing controlled language through a help system integrated into the medical speech translation system (MedSLT). In *MT Summit X*, Phuket, Thaïlande, 2005.

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, septembre 1994.

[Sch02] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, 2002.

[SDV+07] S. Singh, M. Dalal, V. Vachani, P. Bhattacharyya, and O. Damani. Hindi generation from interlingua. In *Machine Translation Summit*, pages 421–428, Copenhague, Danemark, 2007.

[Sei03] Barbara Seidlhofer. Autour du concept d'anglais international : de l'anglais authentique à l'anglais réaliste ? Technical report, Conseil de l'Europe, Strasbourg, 2003.

[Sei04] Barbara Seidlhofer. Research perspectives on teaching english as a lingua franca. *Annual Review of Applied Linguistics*, 24 :209–239, 2004.

[Sei05] Barbara Seidlhofer. *Oxford avdanced learner's dictionary of current English. (7th edition)*, chapter English as a lingua franca. Oxford University Press, Oxford, Royaume-Uni, 2005.

[SKH+00] Furui Sadaoki, Maekawa Kikuo, Isahara Hitoshi, Shinozaki Takahiro, and Ohdaira Takashi. Toward the realization of spontaneous speech recognition −− introduction of a japanese priority program and preliminary results −−. In *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, Pékin, Chine, oct 2000.

[Slo81] D. I. Slobin. *The child's construction of language*, chapter The origins of grammatical encoding of events, pages 187–199. Academic Press, 1981.

[Sot05] Susana Sotillo. Corrective Feedback via Instant Messenger Learning Activities in NS-NNS and NNS-NNS Dyads. *CALICO Journal*, 22(3) :467–496, 2005.

[Sot06] Susana Sotillo. Using instant messaging for collaborative learning : A case study. *Innovate*, février-mars 2006.

[Ste] Jon Stevenson. The language of internet chat rooms. `http://www.demo.inty.net/Units/IRC.htm`.

[SZ05] Mark Seligman and Chengqing Zong. Toward practical spoken language translation. *Machine Translation*, 19/2 :113–137, juin 2005.

[TB06] Tien-Ping Tan and Laurent Besacier. Acoustic model interpolation for non-native speech recognition. In *ICASSP 2006*, Toulouse, 2006.

[Tho08] Alain Thomas. La mesure des progrès lexicaux en FL2. In *Congrès Mondial de Linguistique Française*, Paris, 2008.

[Vau02] Dominique Vaufreydaz. *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue.* PhD thesis, université Joseph Fourier, 2002.

[VBR96] Marie-Thérèse Vasseur, Peter Broeder, and Celia Roberts. *Achieving understanding : discourse in intercultural encounters*, chapter Managing understanding from a minority perspective, pages 65–108. Longman, Londres, Royaume-Uni, 1996.

[VG85] Evangeline Marlos Varonis and Susan M. Gass. Non-native/non-native conversations : a model for negotiation of meaning. *Applied Linguistics*, 6 :71–90, 1985.

[VT04] Hung Vo Trung. SANDOH - un système d'analyse de documents hétérogènes. In *Actes des Journées internationales d'Analyse statistique des Données Textuelles, JADT 2004*, Louvain-la-Neuve, Belgique, mars 2004.

[Wah00] W. Wahlster. *Verbmobil : Foundations of Speech-to-Speech Translation.* Springer-Verlag, Berlin, Allemagne, 2000.

[WCS06] Monika Woszczyna, Paisarn Charoenpornsawat, and Tanja Schultz. Spontaneous thai speech recognition. In *Interspeech 2006 - ICSLP*, 2006.

[WES63] Warren Weaver and Claude Elwood Shannon. *The Mathematical Theory of Communication.* University of Illinois Press, 1963.

[WKL96] Peggy Wong Kim Lai. Comprehension of bitransitive sentences by cantonese-speaking children. Master's thesis, University of Hong Kong, 1996.

[WS03] Zhirong Wang and Tanja Schultz. Non-native spontaneous speech recognition through polyphone decision tree specialization. In *EUROSPEECH 2003*, Genève, Suisse, septembre 2003.

[WW98] E. Wehrli and T. Wehrle. Overview of GBGen. In *Proc. 9th International Workshop on Natural Language Generation*, Niagara-on-the-lake, Canada, 1998.

[Yod01] Benjamin W. Yoder. Spontaneous speech recognition using HMMs. Master's thesis, Massachusetts Institute Of Technology, septembre 2001.

Annexes

1 Tchat multilingue

1.1 Structure de données

Une session de tchat multilingue est, comme une session monolingue, constituée de messages. Et de la même manière, on distinguera des messages « normaux », des commandes, et des notifications d'événements. L'aspect multilingue tient au fait que chaque message pourra comporter plusieurs versions traduites en plus de l'originale. Précisons aussi que l'on compte traiter de la même manière tchat et messagerie instantanée, qui, comme on l'a montré en 1.1, ne diffèrent pas fondamentalement. Enfin, nous souhaitons modéliser certaines informations concernant la configuration des clients, à savoir la langue choisie par l'utilisateur en émission, ainsi que les langues choisies en réception.

Le format de représentation est à nouveau basé sur XML. À la racine `<log>` sont attachés deux types d'éléments : des `<message>`, `<commande>`, et `<evenement>` d'une part, et des éléments `<client>` d'autre part.

Les `<message>`, `<commande>` sont semblables à ceux employés pour le corpus monolingue, à ceci près que le message n'y est plus directement accessible, mais se situe un niveau plus bas, à l'intérieur d'éléments `<original>` et `<traduction>`, respectivement destinés à l'original du message (un message possède toujours un original et un seul) et à ses éventuelles traductions. Les éléments `<original>` et `<traduction>` ne possèdent qu'un attribut `xml:lang`. Le formatage du texte étant possible en XMPP, le message pourra contenir lui-même des balises (en particulier une balise `<body>`). Les `<evenement>` contiennent simplement un code de l'événement décrit.

Les `<client>`, quant à eux, servent à représenter la configuration des clients. Comme l'utilisateur peut changer la configuration de son client au cours d'une conversation, chaque élément `<client>` est caractérisé, outre par un attribut id identifiant le client, par des attributs date et heure, permettant d'associer chaque message avec les configurations clientes correspondantes. Un élément `<client>` peut en outre contenir les éléments suivants (leurs rôles respectifs sont détaillés en 3.5.2) :

- `<pseudo>`, qui donne le pseudonyme de l'utilisateur en PCDATA ;
- `<ressource>`, qui indique le client utilisé, lui aussi en PCDATA ;
- `<languesource>`, dont l'attribut `xml:lang` donne la langue de l'utilisateur ;

- `<languecible>`, dont les attributs `xml:lang` et rang donnent respectivement une langue de traduction pour cet utilisateur, et la préférence accordée à cette langue ;
- `<mode>`, dont l'attribue `value` donne le mode d'affichage.

Ces éléments sont optionnels, mais doivent tous être spécifiés au moins une fois avant l'envoi du premier message. Toutefois, ces informations étant transmises par le serveur au moment de la connexion du client, cette contrainte ne pose pas de difficulté.

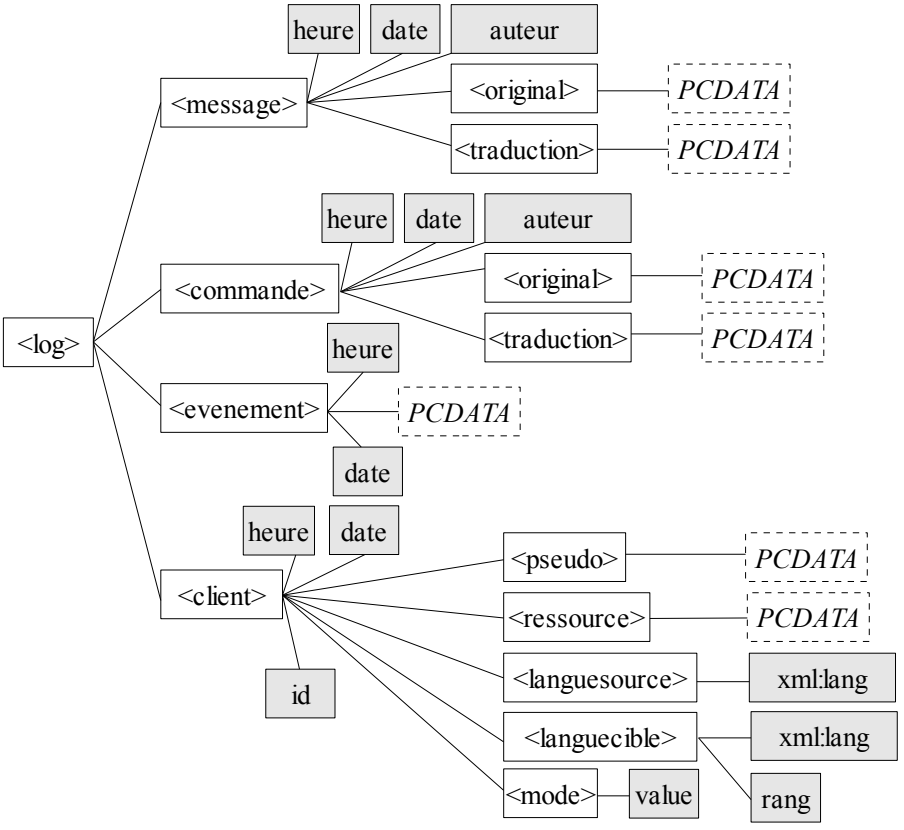


Figure 12 – Format de stockage des sessions de tchat multilingue

1.2 Fonctionnalités

Le système propose trois modes de traduction à chaque utilisateur :

Pas de traduction. Tous les messages sont transférés sans modification.

Sous-titré. Les messages apparaissent deux fois ; une fois sous leur forme originale, et une fois sous forme traduite. Ce mode peut être utile pour les utilisateurs maîtrisant partiellement la langue de leur interlocuteur : la traduction est alors utilisée comme une aide à la compréhension (lexicale notamment).

Version doublée. Les messages originaux ne sont pas visibles, seules les traductions le sont.

Précisons enfin qu'il est possible de changer de mode à n'importe quel moment, sans se déconnecter.

1.3 Configuration

L'utilisation du réseau Jabber nécessite une inscription sur un serveur. Le compte ainsi créé comporte diverses informations, dont certaines sont éditables. C'est notamment le cas du champ « description », qui est généralement laissé vierge. Dans notre système de tchat, on configure son compte client en y déclarant les options choisies. Ces déclarations peuvent intervenir n'importe où dans le champ, éventuellement après la véritable description si l'utilisateur souhaite en saisir une. Quatre types de déclarations sont possibles, correspondant à autant d'options de configuration :

`[lg:xx]` où xx est le code à deux caractères de la langue source de l'utilisateur ;

`[so:x]` où x vaut 1 si les messages originaux (non traduits) doivent être affichés, et 0 sinon ;

`[st:x]` où x vaut 1 si les messages traduits doivent être affichés, et 0 sinon ;

`[lgn:xx]` où xx est un choix de langue cible (de traduction), et n son ordre de préférence. Si le premier choix n'est pas possible étant données les paires de langues disponibles, c'est le second choix qui sera essayé, et ainsi de suite jusqu'à épuisement de toutes les possibilités. Si un message ne peut pas être traduit, il est affiché dans sa langue d'origine.

Par exemple, un utilisateur dont le champ « description » contient « `[lg:fr] [st:1] [so:0] [lg1:fr] [lg2:en]` » est un francophone, qui utilise le mode « version doublée » (pas de message original), et souhaite une traduction de préférence en français, et en anglais si la paire de langue appropriée n'est pas disponible.

Cette méthode est certe peu ergonomique, mais a le mérite de fonctionner avec tous les clients et tous les serveurs. Et ainsi, l'utilisateur peut modifier ces options à n'importe quel moment, sans avoir à se déconnecter.

2 Extraits du corpus de tchat

2.1 Dialogue finalisé

Les pages suivantes présentent un extrait de 500 messages du canal #c++, un canal d'entraide entre développeurs C++, de ce fait plus riche que la moyenne en

dialogues finalisés. L'extrait débute vers 22h, et dure jusqu'au lendemain 18h.

Djoobstil	overdose
zul	faudra m'expliquer l'interet de ce que tu cherche à faire au fait :)
Djoobstil	quel est l'interet d'inventer le C le C++ le python le ruby le TCL le Perl... pkoi pas garder seulement l'ASM ?
Djoobstil	:)
TorF	pourquoi en ASM ? on veut du binaire ! rien que du binaire !
zul	parcxe que si tu codais en asm tu comprendrai que c tres chiant de coder en asm
zul	et le C++ est un langage objet
zul	et les autres sont des langages de scripts
Ossus	et puis pas très portable l'asm
zul	dpnc aucun n'a la meme optique
Djoobstil	bon je vais devoir me dermerder a trouver tt seul
Djoobstil	merci quand même
Djoobstil	:]
Guru_Meditation	moi j'aime bien l'asm c'est ma langue préféré
Guru_Meditation	:)
SeRpenT_705	ciao @demain ++
SysTeM|Failure	» hi Kra|hen
SysTeM|Failure	» re all
Ocean	Je ne sais pas qui est Cahaan.
G-FACTION	-_-'
BotStats15	. : : Les statistiques de ce canal sont disponibles sur : http ://go.botstats.com/ ?c++ : :.
Djoobstil	TorF
BombStrike	grmf c long un ripping de DVD...
Djoobstil	un mot en 0x000... c ça le héxadecimale ?
BombStrike	hum oui je crois...
Djoobstil	k
Ossus	ui 0x c'est de l'hexa
Djoobstil	c magnifique !!!!!!!!!!!!!!!!!!!!!
Djoobstil	sque jveux faire
Djoobstil	c genre une db de Anope :
The_Void	Bn Tlm
Djoobstil	Øð
Djoobstil	des trucs comme ça
Djoobstil	mais j'arrive plus a copier
Djoobstil	:X
G-FACTION[Sleeping]	bn epik ! :)

Djoobstil	zum
Djoobstil	zul
Djoobstil	static unsigned char CRY[26]={'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'} ;
Djoobstil	static unsigned char CRY2[27]=ABCDEFGHIJKLMNOPQRSTUV WXYZ ;
Djoobstil	la seul difference c'est la taille ?
zul	non
zul	le deuxieme est terminé par un \0
Djoobstil	ha ok
zul	donc il s'affichera correctement
Djoobstil	c quoi l'interêt du premier par rapport au 2è ?
zul	aucun :)
zul	ah si se faire chier
Djoobstil	k
Djoobstil	c un rebus du C
Djoobstil	un débris
Djoobstil	:D
TorF	houla c'est pas sur du tout que le dernier caractère soit un '\0'
TorF	j'aurai plutot dit que le caractère est indeterminé
TorF	peut etre je me trompe...
Djoobstil	bon
Djoobstil	moi faut que j'apprenne les héxadécimeaux
Djoobstil	!
zul	Djoobstil : au debut la deuxieme notation n'existait pas en C
zul	TorF : sisi la deuxime assure une null terminated string
Djoobstil	vi
Djoobstil	=)
TorF	ok
Djoobstil	zul tu sais comment anope fait pour ecrire dans un fichier
Djoobstil	comme ça :
zul	bah non je sais pas comme ca :)
Djoobstil	Øð
Djoobstil	des trucs du genre
Djoobstil	Øðÿþ
Djoobstil	
zul	bah comme le reste
Djoobstil	si j'ecris les mots en hexadécimale ils vont apparaitre dans le fichier comme tel
Djoobstil	0x002
Djoobstil	?
zul	ca depend comment tu l'envoie
Djoobstil	en binaire
zul	si tu fais un fprintf(file,"%x",truc) ca passera
zul	mais pkoi tu veux ouvrir ton fichier en binaire

Djoobstil	ché pas
Djoobstil	c stilé
Djoobstil	=)
zul	si c des logs cun fichier normal qu'il faut
Djoobstil	non c un fichier qu'on doit pas pouvoir lire
zul	bah le binaire on peut le lire
Djoobstil	un fichier avec une encryption codé à l'état du binaire encodé héxadécimalement encrypté
Djoobstil	1010111011 01000111
Djoobstil	tu sais lire ça ?
Djoobstil	:X
zul	bah comme ca non
Djoobstil	110100000100111100001111
zul	mais en cherchant ca possible
Djoobstil	bah c bon c des codes
Djoobstil	c juste que se soit pas tt de suite visible quoi
Djoobstil	pr pas que les noobs puissent le lire kwa
zul	disons qu'en supposant que c un texte vais etre amené a lire les bits 4 par 4
Djoobstil	101101001111001110011100 0
Djoobstil	110100000100111100001111
Djoobstil	100001111111
Djoobstil	:X
Djoobstil	10010110111010011
TorF	tu peux utiliser le xor pour crypter du texte facilement
Djoobstil	10111011000100101100001011 0011
TorF	A xor X =&gt ; Y Y xor Y =&gt ; A
Djoobstil	10000000011101111110011001000 10111
Djoobstil	:X
zul	je connais xor :)
Djoobstil	BONJP0R
Djoobstil	:DDDDDDDDDDDD
Djoobstil	EE9D0CD
Djoobstil	^^
Djoobstil	HOQ93JJ
Djoobstil	2C5B1M8A
Djoobstil	6EE3ADG7
Djoobstil	8621365527
Djoobstil	18073401B3
Djoobstil	A751AC828
Djoobstil	ça c jolie
Djoobstil	=)
Djoobstil	119AC
TorF	sinon tu reprends une lib de cryptage...

Djoobstil	t'en as une ?
TorF	non mais ca doit se trouver facilement
TorF	tu peux zipper en utilisant un mot de passe par exemple
Djoobstil	static unsigned long crc32_tab[] = { 0x00000000L, 0x77073096L, 0xee0e612cL, 0x990951baL, 0x076dc419L, 0x706af48fL, 0xe963a535L, 0x9e6495a3L, 0x0edb8832L, 0x79dcb8a4L, 0xe0d5e91eL, 0x97d2d988L, 0x09b64c2bL, 0x7eb17cbdL, 0xe7b82d07L,
Djoobstil	y en a 54lignes de ça
Djoobstil	comment ils encryptent a partir de ça ?
Djoobstil	jpense pas au rand()
zul	c du md5
zul	mais je connais pas l'algo exacte
Djoobstil	unsigned long our_crc32(const unsigned char *s, unsigned int len)
Djoobstil	{
Djoobstil	unsigned int i ;
Djoobstil	unsigned long crc32val ;
Djoobstil	crc32val = 0 ;
Djoobstil	for (i = 0 ; i < len ; i ++)
Djoobstil	{
Djoobstil	crc32val =
Djoobstil	crc32_tab[(crc32val ^ s[i]) & 0xff] ^
Djoobstil	(crc32val » 8) ;
Djoobstil	}
Djoobstil	return crc32val ;
Djoobstil	}
TorF	http ://www.pgpi.org/products/pgp/versions/freeware/winxp/8.0/
TorF	pour crypter en PGP (interdit en france)
Djoobstil	lol
zul	c interdit en fr bah merde alors :)
Djoobstil	j'avais PGP
Djoobstil	le programme
Djoobstil	;)
zul	tu es sur TorF que c encore interdit
TorF	non
TorF	j'ai vu ca sur un site... s'il date de 50 ans
Djoobstil	pkoi interdit ?
TorF	d'ailleurs c'est chiant les gars qui donnent JAMAIS la date de leur site...
TorF	parce que trop difficilement décryptable
zul	a mon avis c plus interdit ils ont assoupli la taille des clé utilisable
Djoobstil	lol
Djoobstil	et alors ça fait quoi que se soit pas décryptable ?
Djoobstil	c fait pour ça non ? :D
zul	bah pour toi oui mais pour les services secrets non :)

Djoobstil	un header ça sert a mettre les fonctions ?
Djoobstil	définir*
zul	declarer
TorF	bye
Ocean	[zul] It doesn't work until it's right ! !
Djoobstil	putain
Djoobstil	erf
Djoobstil	services de merde
Djoobstil	réso de merde
Djoobstil	:(
zul	lol
zul	un serveur a splitter c la vie
Djoobstil	vie de merde
Djoobstil	:(
Djoobstil	merci
Djoobstil	=)
Djoobstil	t'y a cru hein
Djoobstil	:D
Djoobstil	zul t'as été vexé là dernière fois qu'on a parlé des OS ?
zul	bah je peux pas supporter les gens qui sont pas capable de lire les documents qu'on leur fournit
Djoobstil	oué
Djoobstil	moi je lis rien
Djoobstil	=)
Djoobstil	qu'on me fournisse ou pas
Djoobstil	le dernier livre que j'ai lu c t en CM1
Djoobstil	là je suis en 3è
Djoobstil	et encore j'en avais lu le quart
zul	ca aide de lire
Djoobstil	ha bon
Djoobstil	(=
zul	sisi y a pleins de documentation
zul	y a pas que des livres chaints
zul	meme si il y a de tres bon livre
Djoobstil	faut utiliser quoi pour chopper si l'user appui sur ctrl + F1 ?
zul	Djoobstil : je sais pas comment hooker deux touches
BotStats15	. : : Les statistiques de ce canal sont disponibles sur : http ://go.botstats.com/ ?c++ : :.
G-BOT01	Je suis activé sur #c++. Tapez !commandes01 pour mes commandes.
BombStrike	re
camje_lemon	Yo
BombStrike	yo

camje_lemon	ReP :/
BombStrike	rep :)
camje_lemon	^^
camje_lemon	loul
camje_lemon	=)
camje_lemon	jV manger @tte de suite
BombStrike	bon miam :)
camje_lemon	«+BombStrike» merci et rep
camje_lemon	«Ossus» Yeah !
camje_lemon	yo
Ocean	[zul] It doesn't work until it's right !!
BombStrike	pouet :)
BotStats15	. : : Les statistiques de ce canal sont disponibles sur : http ://go.botstats.com/ ?c++ : :.
The_Void	helloww tlm
Djoobstil	zul =)
zul	oui Djoobstil
Djoobstil	lut
Djoobstil	:))
zul	kikoo
Djoobstil	bon
Djoobstil	ça me vnr
Djoobstil	jpeux pas compiler asuka sous debian
zul	c surement marqué pkoi
Djoobstil	oué
Djoobstil	bash-2.05b$ ./configure
Djoobstil	loading cache ./config.cache
Djoobstil	checking for installation prefix...
Djoobstil	checking host system type... ./config.guess : unable to guess system type
Djoobstil	This script, last modified 2002-03-04, has failed to recognize
Djoobstil	the operating system you are using. It is advised that you
Djoobstil	download the most up to date version of the config scripts from
Djoobstil	ftp ://ftp.gnu.org/pub/gnu/config/
Djoobstil	If the version you run (./config.guess) is already up to date, please
Djoobstil	send the following data and any information you think might be
Djoobstil	pertinent to &lt ;config-patches@gnu.org&gt ; in order to provide the needed
Djoobstil	information to handle your system.
Djoobstil	zul
zul	oui
Djoobstil	Solaris pour x86 c'est vraiement pas bien ? :/
zul	Djoobstil : ca sux solaris c pas libre

Djoobstil	20€... moi jtrouve ça libre
zul	Djoobstil : pas gratuit libre
zul	grande différence
Djoobstil	c quoi libre ?
zul	libre ca veut dire que n'importe qui peut avoir acces au source code et le modifer cad n'importe qui peut l'ameliorer
Djoobstil	wé
Djoobstil	moi jveux pas l'ameliorer
Djoobstil	il est déjà très bien
Djoobstil	enfin, si on a plus de 12processeurs :X
zul	Djoobstil : bah vi mais bon ce qui est pas libre suxe par definition
Djoobstil	par ta definition
Djoobstil	:)
zul	par definition universelle
zul	billou a dit : 'si on avait inventé les brevets il y a 50 ans, je ne serai surement pas patron de la plus entreprise de soft du monde car l'ordinateur n'existerai meme pas'
SysTeM|Failure	» re ici
Djoobstil	la plus entreprise ?
zul	+grande
Djoobstil	:)
SkyMaster	Lu !
SysTeM|Failure[afk]	-Away- —&gt ;Raison » réparation, dépense, montage ! ! PC chibré ! —&gt ;Départ à » 13 :12 :56 ! SunScript
TorF	salut
G-BOT01	Je suis activé sur #c++. Tapez !commandes01 pour mes commandes.
BombStrike	winamp (playing) K̃almah - Cloned Insanity (04 :13/128kbps)
BombStrike	winamp (playing) K̃almah - The Third, The Magical (05 :28/128kbps)
Djoobstil	TorF
TorF	c'est moi
Djoobstil	comment tu fais pour transformer un mot en héxa ?
Djoobstil	au fait
Djoobstil	l'héxa
Djoobstil	c pas les adresses des variables ?
TorF	non
TorF	l'hexa c'est comme le décimal ou le binaire
TorF	c'est une facon de représenter un nombre
TorF	tu peux avoir l'adresse d'une variable en binaire, en décimal, en octal ou en ce que tu veux
TorF	et t'appelles quoi un mot ? une chaine de caractères ?
Djoobstil	oui
Djoobstil	dans mon livre de C++
Djoobstil	ils disent
Djoobstil	qu'une var peut se trouver à 0x123
Djoobstil	pour expliquer &

Djoobstil	pour expliquer & et *
Djoobstil	:)
TorF	par convention on utilise l'hexa pour les adresses oui
Djoobstil	k
TorF	et bien pour transformer un mot en hexa
TorF	tu fais une fonction qui affiche un caractère en hexa
TorF	et tu appelles cette fonction avec chaque caractère de ton mot
Djoobstil	oui
Djoobstil	avec une while
Djoobstil	mais comment on transforme un caractère en héxa ?
Ocean	[Cahaan] &lt ;DH-Team&gt ;
TorF	pour l'afficher en hexa ?
Ocean	[zul] It doesn't work until it's right ! !
The_Void	vais manger ++
TorF	je te copie le code en pv
The_Void	Re
op-hium	ne pas ouvrir les url de ce type http ://won.attaq.***/lol.jpg et http ://www.geocities.***/privpics2004/ RISQUE DE VIRUS
Djoobstil	lol
Djoobstil	ho Cahaan =)
The_Void	Re
Got3n	re
Kra|hen[Aw]	reuh les G-FACTION[Velo]
Kra|hen[Aw]	les gars je voulai dire
Kra|hen[Aw]	LOL
Kra|hen[Aw]	Cahaan
Kra|hen[Aw]	:D
Kra|hen	salu SeRpenT_705
SysTeM|Failure[oQp]	-Away- —&gt ;Nouvelle Raison » montage PC ! —&gt ; SunScript
The_Void	a++
Kra|hen	Cahaan
Kra|hen	:D
Kra|hen	qd t la
Kra|hen	dis le mwa
Kra|hen	pliuz
Djoobstil	faut qu'il me donne mon access :p
Kra|hen	tu lui demanderas
Kra|hen	=)
Djoobstil	oé
Djoobstil	tien un voice
Djoobstil	parceque t gentil

Djoobstil	:p
Kra|hen	MERCHIIIIIIIIIII
Kra|hen	=))))
Djoobstil	ouéééééé
The_Void	re
Djoobstil	TorF c un pGm
Kra|hen[Aw]	pGm c koi ?
SkyMaster	Lu !
Djoobstil	a la base
Djoobstil	c'est pro gamer
Djoobstil	et après ça a dérivé en pro
Djoobstil	un gars qui est un pro tu dis c un pGm
TorF	tout est relatif comme on dit hein
Djoobstil	rheu
Djoobstil	zul
Djoobstil	op plz
Djoobstil	=)
FireDragoon	alu
Djoobstil	thx
Djoobstil	:))
Kra|hen[Aw]	Cahaan
Kra|hen[Aw]	Cahaannnnnnnnnn
Kra|hen[Aw]	Cahaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaannnnnnnnnnn
Kra|hen[Aw]	mdr ;)
Kra|hen[Aw]	il est jamais la
Kra|hen[Aw]	:p
zul	ouep c embetant
zul	il doit etre en partiel
TorF	il est parti pieuter
Ocean	Cahaan est sur le canal actuellement !
zul	k :)
BotStats15	. : : Les statistiques de ce canal sont disponibles sur : http ://go.botstats.com/ ?c++ : :.
Djoobstil	lol zul
Djoobstil	[17 :53] &lt ;zul&gt ; !seen cahaan
Djoobstil	[17 :53] &lt ;Ocean&gt ; Cahaan est sur le canal actuellement !
Djoobstil	[17 :53] &lt ;zul&gt ; k :)
zul	Djoobstil : g pas la liste des users qui s'affichent sur irssi
zul	fo que je te tape /who
Djoobstil	lol
Djoobstil	c trop nul

Djoobstil	:
Djoobstil	:D
Djoobstil	mIRC pawa
Djoobstil	franchement c un bon client
Djoobstil	jvois pas pkoi y a des gens qui dient mIRC SuX
zul	tu as payé ton mirc j'espere
Djoobstil	non
Djoobstil	c pas obligatoire
zul	bah si
Djoobstil	sinon il ferait un blockage
zul	c obligatoir
Djoobstil	...
Djoobstil	bah
Djoobstil	il a pas fait de blockage
Djoobstil	c que c pas obligé alors
zul	Djoobstil : bah il perd pas son temps coder un truc inutile :)
Djoobstil	sinon il en aurait mit un
zul	y aurait une protection tu l'aurai cracké
Djoobstil	a mon avi il code au moin 100x mieu que toi
Djoobstil	donc il fait ça en 1mn même pas
TorF	je me demande pourquoi Djoobstil est op, c'est une honte !
TorF	lol
zul	bah et le crack sera trouvé en 5 min ca sert a rien
Djoobstil	BombStrike y avait combien de lignes de code et de pages dans mIRC ?
zul	vi c clair
Djoobstil	jsuis op parceque j'étais sur ce canal avant vous
Kra|hen[Aw]	:D
zul	Djoobstil : de toute facon fo le payer mirc , il est gratuit pour 30 jours
zul	alors si tu le trouve bien tu le paye au moins
zul	ca me parait le minimum
Djoobstil	j'ai pas les sous
Djoobstil	j'ai même pas les sous de réparer mon 2è ordi
zul	Djoobstil : bah tu l'utilsie pas alors
Kra|hen[Aw]	mais c un access en fonction des faculté au dev c++
Djoobstil	..
zul	Djoobstil : et si tu trouvais ca bien tu economiserai pour l'acheter
Djoobstil	zul dis ça au millions de personnes qui l'utilisent sans le payer
Djoobstil	:)
zul	Djoobstil : bah je le fais
zul	et la différence c que toi tu dis que c genial

Djoobstil	non
zul	alors que certains ne connaissent que lui ce qui different
Djoobstil	je dis que c bien
Djoobstil	c different !
TorF	il y a des gens honetes, et d'autres malhonetes
Djoobstil	bon tu vas me laisser finir de manger mon Baton de berger Justin Bridou merde ! :p
zul	tu fais partie des gens malhonnetes :)
Djoobstil	.
Djoobstil	et alors
Djoobstil	menfoo d'être malhonnete
Djoobstil	c'est ça l'internet
Djoobstil	c'est ça la vie
Djoobstil	:)
Djoobstil	mais quesqui fait
zul	j'aime vraiment cette putin de mentalité de merde
Djoobstil	:D
Djoobstil	ben c bien
Djoobstil	lol
zul	l'internet ca a ete cree pour que la connaissance se partage
zul	pas pour etre malhonnete
Djoobstil	jcritique pas ta mentalitée
zul	je remercie tous les gens comme toi
Djoobstil	Cahaan disait, pour que je sois op il faut que je respecte les autres
Djoobstil	questu fous op zul ?
Djoobstil	respecte mes pensés
Djoobstil	j'en fais ce que je veux
Djoobstil	c'est un droit.
zul	Djoobstil : je n ai pas a respecter ca
zul	si cahaan veut me deoper qu'il le fasse
TorF	non, on a pas le droit d'être malhonette lol
Djoobstil	vas lire ça merci :)
Djoobstil	G-FACTION GRENOBLOIIIIS
G-FACTION	GRENOBLE EN FORCE
G-FACTION	hello Cahaan =)
Djoobstil	klr
zul	Djoobstil : Djoobstil tu viole pourtant un certain nb de regles du lien
Djoobstil	lequels ?
Djoobstil	pas payer un sharewar ?
Djoobstil	ça c'est la cnil
zul	bah vi c du vol :)

Djoobstil	pas les droits de l'homme
Djoobstil	hein
Djoobstil	:)
zul	Djoobstil : c du vol
zul	rien a prouver de plus
Djoobstil	bon
Djoobstil	————————————-(Pause Baton de Berger Justin Bridou)————————————-
zul	j'aime pas les gens malhonnetes
TorF	utiliser les droits de l'homme pour justifier sa malhonneteté... c'est à mdr =)
Djoobstil	c pas mon probleme =)
Djoobstil	lol
Djoobstil	klr TorF
Djoobstil	:D
zul	je sais pas ce qui me retient de t'ignore
zul	surtout qu'il l'a jaamsi lu
Djoobstil	parceque tu m'aime
Djoobstil	si je l'ai lu l'an dernier
Djoobstil	j'ai même eu 2tests dessus
Djoobstil	:)
zul	damn tu la connais par coeur
Djoobstil	quoi ?
Djoobstil	la constitution ?
Djoobstil	la loie ?
Djoobstil	ou les droits de l'homme
zul	la declaration des droits de l'homme et du citoten
Djoobstil	?
Djoobstil	non
Djoobstil	et toi ?
Djoobstil	mais je connais
zul	non :)
Djoobstil	-Libertée de penser
Djoobstil	:))
zul	Djoobstil : bien sur :)
zul	mais on a le droit de repudier certaine chose
zul	et ta liberté s'arrete ou celles des autres commencent
Djoobstil	wé
Djoobstil	ta libertée s'arrête là ou la mienne commence
Djoobstil	essaye de pas en prendre trop
Djoobstil	tu risque de prendre de la mienne
Djoobstil	:)

Djoobstil	franchement
Djoobstil	m'embrouiller avec zul, c'est 1000000000000000000x mieu que avec aerith
Djoobstil	parcequ'avec zul ça sert a quelque chose
Djoobstil	on parle des droits
zul	non
Djoobstil	et tout ça
zul	tu change pas
zul	c inutile
TorF	haaaaa un bon coup de troll :)
Djoobstil	aerith elle c'est connard bouffon apprends a scripter...
Djoobstil	le truc trop lucratif
Djoobstil	:D
zul	je reste poli tjs
Djoobstil	héhé tant que tu le reste, je le reste
zul	si un mec me soule je l'ignore ou je me casse
Djoobstil	moi jme casse pas
Djoobstil	c'est ue question d'étique personnel :)
Djoobstil	c'est une question d'étique personnel :)
zul	Djoobstil : bah la fierté stupide n'amene a rien
zul	et je vois pas comment on peut parler d'etique personnel qd on a que des logiciels pirates

2.2 Dialogue non finalisé

Les pages suivantes présentent un extrait de 500 messages du canal #18-25ans, un canal de discussions variées, dont les dialogues présentent généralement un caractère moins finalisé que #c++. L'extrait correspond à une heure de discussion, entre 12h et 13h.

Avrilav	fabolous : quoi ?
Avrilav	ARF
fabolous	lol
Avrilav	tu te prend pour johnny...*
Avrilav	lol
fabolous	a que oui ma jolie
Avrilav	latatia
PitiRom	lollllll
PitiRom	a non ppda
PitiRom	que ma femme c pas laetitia
PitiRom	cest la tia tia
Avrilav	elle est con latatia je trouve, elle fait naise
Avrilav	je pensais pas pourtant
PitiRom	lol
Avrilav	lol
PitiRom	im so naked
PitiRom	around you
Mike_cours	re ici
Gavaway	bonjour San-A
San-A	lut Gavaway
cosette_lycee	kookoo tlm ! !
PitiRom	salut cosette_lycee
cosette_lycee	j'ai fini les cours la pour la journee yeahhh
cosette_lycee	chui contente ! ! !tlm sen fiche mais chui cntente ! ! !
Gavaway	je suis content pour toi
cosette_lycee	merci gavaway
PitiRom	bah non cosette_lycee
PitiRom	je partage ta joie
Gavaway	cosette_lycee : tu as plus qu'à passer à tes devoirs
cosette_lycee	mechant gavaway ! ! !
cosette_lycee	j'en sors du lycee la molo !
cosette_lycee	vous faites koi la ? ?
PitiRom	rien
PitiRom	jai la flemme de bouger
Gavaway	la ?
Gavaway	rien
Gavaway	je me demande quel train je vaois prendre pour rentrer chez moi
cosette_lycee	tu es de ou ?
cosette_lycee	gavaway
Gavaway	Le Mans

Gavaway	et toi ?
cosette_lycee	83
Gavaway	c'est ou ca ?
cosette_lycee	le sud
Gavaway	le var ?
cosette_lycee	la ou y a tt plens de beaux garcons
kotay	lolilou la bonne blague :p
icarius	salut
Gavaway	cosette_lycee : tu mérites les beaux garcons toi ?
cosette_lycee	oui ! ! !
San-A	ou pas
San-A	(ou photos :))
San-A	histoire qu'on voit
cosette_lycee	quoi donc san antonio ?
icarius	qui connait Ragnarok online ? ? ?
kotay	lool :) San-A le pervers :p
Gavaway	moi.
San-A	kotay : moi ? du tout
Gavaway	cosette_lycee : envoie la photo
Gavaway	qu'on juge ca par nous meme
San-A	cosette_lycee : tu connais San-Antonio toi ? :)
San-A	c rare de nos jours
icarius	personne ? ?
cosette_lycee	sis ps idiote a spoint
San-A	bah y en a qui connaissent pas
cosette_lycee	belle et intelligente c tt moi ca ! !
San-A	intelligente ? prouve le :)
icarius	personne :-/
San-A	icarius : personne
kotay	lolilou :)
cosette_lycee	je ris
San-A	et ?
kotay	et t'es blonde tant que t'y es cosette_lycee ; nan ? :D
cosette_lycee	ah non pas blonde ! !
San-A	une blonde intelligente ?
San-A	wouaouh !
San-A	ah non
cosette_lycee	j'ai les cheveux bleus
San-A	brune ? chatain ?
San-A	magnifique couleur naturelle !

cosette_lycee	bleue ! !
cosette_lycee	c beau hein
San-A	bof
cosette_lycee	oh
San-A	ça dépend du bleu
Avrilav	ha vive la weed
San-A	:)
cosette_lycee	un bleu elctrike
San-A	la quoi Avrilav ?
Avrilav	weed
San-A	cosette_lycee : càd ?
San-A	Avrilav : c quoi c'te chose ?
cosette_lycee	suis keuponne !
Avrilav	San-A : arf traduit
cosette_lycee	ptdr ! ! !
San-A	-_-
San-A	o&lt ; o&lt ; o&lt ;
San-A	j'en sais rien moi
kotay	vive les sctroumpfette :D
cosette_lycee	ktay ! ! ui ma pool
San-A	kotay : la schtroumfette est blonde !
cosette_lycee	c une fausse blonde !
cosette_lycee	san a t'es comment ?
Mike	blond cendré !
cosette_lycee	pouahhh !
cosette_lycee	kel horreur !
cosette_lycee	j'aimeke les bruns ! !
Mike	pas grave ; tu ne me vois pas
Gavaway	Mike : parce que SAN A ca vient de SAN ANTONIO ?
Mike	Gavaway : oui
Gavaway	je pensais que ca venait d'un manga ridicule encor
Mike	non non
cosette_lycee	y a t'il des beaux mecs bruns ici ?
Gavaway	cosette_lycee : oui moi
Mike	...
cosette_lycee	ah ?
Gavaway	et oui
cosette_lycee	fais voir ! !
Mike	pas vrai
kotay	... lolilou :)

Gavaway	mais tu es un peu jeune pour moi ma chérie.
Gavaway	Mike : toi tu m'as pas vu récemment :)
cosette_lycee	hey oh !!
Gavaway	(un pari perdu et hop me voila brun)
Gavaway	cosette_lycee : tu as quel age ?
Mike	mouarf
cosette_lycee	18
cosette_lycee	et ti ?
Mike	terminale ?
Gavaway	18 ans ... hmmm interressant
Gavaway	21
cosette_lycee	terminale vi
Mike	L ?
cosette_lycee	ah ca peut le faire la nan ?
cosette_lycee	pas l non en eco
Gavaway	cosette_lycee : tu es quand meme loin.
Gavaway	(vive eco)
Mike	et tu te dis intelligente ? ? ? o_O
Mike	c nul l'éco
cosette_lycee	oui mike !!!
Mike	eh bah
Mike	ça fait peur
Mike	tu vises HEC ?
cosette_lycee	ouinnnn fais gaffe !
Mike	à quoi ?
cosette_lycee	nn l'ecole centrale a paris
Gavaway	(nature)
Gavaway	c'est bien l'école centrale
Mike	c une école scientifique
cosette_lycee	je sais !!
Gavaway	mais je te conseille plutot le parcours standard : science po Paris et Ena
cosette_lycee	je reve ..
Mike	bah c pas en faisant éco qu'on peut y entrer, si ?
cosette_lycee	laisse moi rever
Gavaway	Mike : cest un concours...
Gavaway	suffit de bucher
cosette_lycee	j'aime pas bosser
Mike	merci
kotay	rahlala... vive la fac, y a rien de mieux pour glander :p
cosette_lycee	j'ai mieux a faire

Mike	tu connais le concours centrale-supelec Gavaway ? ? ?
Gavaway	Mike : ben non.
cosette_lycee	je vais manger
Gavaway	mais comme tout concours suffit de bucher
Gavaway	j'ai pas dit que c'était simple.
Mike	bon app cosette_lycee
cosette_lycee	kiss a tous
Gavaway	cosette_lycee : tu reviendras ?
cosette_lycee	tt a l'heure oui
Gavaway	coool
cosette_lycee	tu seras la ?
Gavaway	possible
cosette_lycee	bah oui ou non
Gavaway	ca depend de l'heure ou tu reviens.
Mike	logique
Mike	:)
Avrilav	tu as volé as volé l'orange
Mike	Avrilav : ? ? ? ?
cosette_lycee	et bien je reviens dans un quart dheure ?
Gavaway	ok
Gavaway	je serais la
cosette_lycee	chouette
Gavaway	Avrilav : toi tu écoute n'importe quoi
Mike	Arthurb : c un brun ténébreux intelligent
Mike	spécialement pour cosette_lycee :p
Avrilav	Gavaway : non c'est la radio..lol
Gavaway	Mike : elle est ou lila ?
Avrilav	une nuit sur son épauuuuuuule
Mike	comment ça elle est où ?
Mike	où elle habite ?
Gavaway	en ce moment, ou k'elle est ?
Tsubasa_C	yop tlm :)
Mike	j'en sais rien moi
Avrilav	il me faut qqch de plus fort
Gavaway	moi je kiffe souchon !
beaugoss	bjr tous le mondes
Avrilav	Gavaway : halala
Mike	ACTION écoute Makina - Can can
Gavaway	tu aimes pas ?
beaugoss	pas de miss ou quoi ?

Mike	c nul souchon ! ! !
Avrilav	Gavaway : bah la j'écoute micropoint donc a coté souchon c'est de la gnognotte...
beaugoss	et voulsz
Gavaway	beaugoss : non c'est un chan gay ici
Mike	connais pas micropoint
Mike	Gavaway : stoi le gay
Gavaway	Avrilav : c'est quel genre micropoint ?
Gavaway	Mike : toi on en reparlera quand tu aurais touché un vagin.
Avrilav	Mike : dl, tu trouveras facilement sur kazaa
beaugoss	serieux ?
Mike	j'ai pas kazza
Mike	kazaa
Mike	marche pas
Gavaway	beaugoss : ben oui
Avrilav	Gavaway : hardcore trés trés bourrin
Gavaway	ah ouais c'estpas le meme style...
Avrilav	Gavaway : spa pour les tafioles quoi
Avrilav	lol
Mike	ACTION écoute BlaBla(Hardcore ver)
Avrilav	Gavaway : tu as kazaa ou truc dans le meme genre ?
Gavaway	non
Avrilav	ARF
Gavaway	le piratage c'est mal.
Avrilav	lol
Mike	ou pas :)
Avrilav	ha mais g envie de danser , d'aller en teuf la
Gavaway	Avrilav : tu devrais écouter souchon quand meme
jacques1	salut
Mike	c de la merde souchon ...
Avrilav	Gavaway : je connais t'inquiete mais ça me fait plus triper que ça
Gavaway	Mike : souchon pour draguer la petite bourgeoise, c'est tres bien.
Mike	m'en fous de la petite bourgeoise !
kickuchi	lu
Gavaway	Mike : pourtant c'est bon la petite bourgeoise
Gavaway	ca craque sous la dent.
Mike	...
Mike	mouaif
cl54	salut
Mike	faut encore supporter souchon
Mike	lut kickuchi et cl54

Mike	ACTION écoute Vivamine RMX par MKN
Gavaway	Mike : moi ca m'étonne pas que tu aies des probleme avec les femmes si tu leur fait écouter ton truc de sauvage quand tu les prend dans tes bras
Mike	euh
Mike	j'en ai jamais pris une dans mes bras !
Avrilav	le theatre du bruit est notre puissance
Gavaway	Mike : dommage
Mike	et je supporte pas les slows (me faudrait des boules quiès)
Gavaway	Mike : trouve toi une femme ici
Avrilav	Gavaway : tu plaisantes
Gavaway	Avrilav (si tu es une fille) tu veux pas partager un peu de temps avec Mike ? il aime la musique de bourrin.
Mike	t'inquiète pas pour moi, je suis toujours sur mon coup
cl54	y a t'il un beau mec dans le coin
Mike	Avrilav est une fille
Avrilav	Mike : spa dla musique de bourrin...
Gavaway	cl54 : oui y a Mike
Mike	cl54 : ouais, y a Gavaway
Mike	lol
cl54	ok les gars
Mike	il a tapé plus vite que moi
Gavaway	Mike : nancy c'est ton coin.
cl54	alors qu'est ce que vousracontez
Gavaway	cl54 : j'essaye de trouver une copine à Mike
Gavaway	mais c'est dur
Mike	c un trou paumé Nancy, c encore pire que Dijon
Gavaway	il est difficile
cl54	il recherche quoi
Gavaway	cl54 : n'importe quoi
Avrilav	dijon ça craint oui
Gavaway	tant qu'il y a un vagin.
Gavaway	mais non dijon c'est rigolo
Gavaway	le TGV qui fait pARIS DOLE s'y arrete
Meublo_Crookah	yo Avrilav
Meublo_Crookah	ou yo hoe
Avrilav	coucou Meublo_Crookah
cl54	ouais donc c paune cop qu'il cherche mais un trou
Mike	ACTION écoute Trance_Energy_2001_-_Volume_1_-_06_-_POTATOHEADS_-_Mix_The_M
Meublo_Crookah	:)
Meublo_Crookah	sava ? ?
Gavaway	cl54 : non je rigole
Gavaway	il cherche une copine.

Avrilav	Meublo_Crookah : oué
Mike	...
Avrilav	michalak parti... ?
Mike	chuis sur un coup, merci du coup de main
Avrilav	ha Tsubasa_C[DivX]
cl54	ok c cool et toi tu cherche quoi
Gavaway	cl54 : alors ASV !
icarius	personne ne conait ragnarok ? ?
Avrilav	si
Mike	on le connait déjà son asv
Gavaway	icarius : si
Gavaway	c'est un MMORPG à la sauce japoniaise
Gavaway	pourquoi ?
icarius	ah ca vous dirais d'y jouer ?,
Gavaway	non
Gavaway	j'aime pas ce genre de jeu.
icarius	ah
Mike	c quoi comme type de jeu ?
Gavaway	moi je veux une hache +5 mon armure en poil de dragon +3 et des pouvoirs magique
icarius	c'est un RAP
icarius	RPG
Mike	oki
Gavaway	icarius : oui oui
Gavaway	je sais.
Gavaway	mais la realisation laisse àd ésirer
Gavaway	lm'abonnement coute combien ?
icarius	10eruo
Gavaway	CHER !
Gavaway	10 € / mois ?
icarius	oui
icarius	Qq'un aime les RPG ici ? ?
Gavaway	oui moi
Meublo_Crookah	bon app tlm
Gavaway	icarius : mais surtout les solos
Avrilav	ha ouai j'vé allé bouffé
Gavaway	je suis un fan de morrowind et des baldurs gate
Gavaway	et des fallout
Gavaway	bon app Avrilav
Meublo_Crookah	bon app Avrilav alors
Mike	bon app Avrilav

Avrilav	bon app
Gavaway	icarius : et toi tu aimes quoi ?
Meublo_Crookah	merci a vous
icarius	moi je joue a ragnarok
Alvyss-Sanae	Ragnarok Online, lol.
Gavaway	et a part ca ?
Alvyss-Sanae	Ils n'ont pas encore fait faillite ? :)
Gavaway	Alvyss-Sanae : faut croire que non
Kirby	re
icarius	hum a plein d'autre jeux CS diablo DAOC etc..
Gavaway	à 10 € / Mois ca doit meme etre rentable.
Mike	c quoi DAOC ?
Meublo_Crookah	c trop echec kom jeu, en plus il faut payer
Gavaway	Dark age of camelot
Gavaway	Alvyss-Sanae : tu joues pas à ce genre de chose ?
Alvyss-Sanae	Gravity sont pourtant de vrais professionels pour agacer les joueurs, backrolls intempestifs, entretien , erreur de la part des GM , Ban par erreur ...
Alvyss-Sanae	Non, mais je connais bien.
icarius	si qq'un aime les RPG j'ai un bon plan pour eux
Alvyss-Sanae	Je suis toute ouïe.
Gavaway	icarius : j'iame pas les RPG online.
Avrilav	houla
Avrilav	Mike
Mike	oui ?
Gavaway	Alvyss-Sanae : je sais pas moi... j'ai jamais essayé ce jeu... mais avant de sortir il était deja dépassé techniquement
Mike	avec toi tout est dépassé techniquement
Alvyss-Sanae	Tout n'est pas technique.
Avrilav	Mike : dl plutot init data c'est plus représentatif du groupe
icarius	laisser moi un message et je vous dit c quoi le plan (c pas un plan foireux)
Gavaway	Mike : mais non.
Alvyss-Sanae	Abrège icarius.
Gavaway	Alvyss-Sanae : en effet mais à 10 € / mois on peut demander qqc de plus joli
Alvyss-Sanae	Où ne pas payer.
Alvyss-Sanae	Ou*
Melanie|pula	a toute all
icarius	Alvyss-Sanae oui je sait ou tu peux jouer a ragna sans payer
Gavaway	moi je suis un inconditionnel de Fallout.... c'est tout vieux, dans une rezolution du moyen age... c'est dépassé mais COMMENT C EST BON !
Alvyss-Sanae	Oui, sur des servers privés icarius.
icarius	oui
Alvyss-Sanae	Mais quasiment aucun intérêt ...
Gavaway	(y a personne sur les serveurs privés)

icarius	ah et pourquoi ?
Alvyss-Sanae	L'intérêt d'un jeu online, c'est d'y jouer avec énormément de monde.
Mike	pas forcément énormément
Alvyss-Sanae	La plupart du temps, ces servers sont vides.
icarius	ben si y a personne qui vien forcement q'il y a personne
Gavaway	icarius : c'est un peu comme faire une crouse de voiture tout seul
Gavaway	c'est un peu à chier.
Alvyss-Sanae	Aucune compétition.
icarius	mouai
Gavaway	(comment ca TUE souchon)
Alvyss-Sanae	Tu lvl ton personage pour aller tuer 2 pixels alignés.
Gavaway	icarius : attends sagement everquest 2 !
Gavaway	comme tout le monde.
Mike	ACTION écoute Mr Joy - Everybody Say Hou !.. Hou !.. Hou ! 2003
Gavaway	moi j'écoute ca : K :\MP3\Alain.Souchon\Alain Souchon\J'veux du live\17 - Le baiser.mp3
Gavaway	et ca rox !
kickuchi	a+++
Mike	Gavaway : ça suxx
Avrilav	Gavaway : tu t'éclates vraiment la ?
Gavaway	Avrilav : ben j'adore souchon
Gavaway	c'est tout.
Mike	il est fou
Gavaway	j'adore les chansons à texte
Mike	stout
Avrilav	ouai c'est fou
Mike	ACTION écoute Jan Wayne - Because The Night
Gavaway	K :\MP3\Alain.Souchon\Alain Souchon\J'veux du live\12 - Foule sentimentale.mp3
Gavaway	ca aussi c'est super !
Meublo_Crookah	lick me baby i know u like it
Meublo_Crookah	yeh
Mike	ou pas
Gavaway	Avrilav : nan tu vois tu serais un peu plus ouverte je te ferais écouter.
icarius	bon allez je vous laisse
Gavaway	bye irelandj
Gavaway	seul
cosette_lycee	voila
Gavaway	re cosette_lycee
Avrilav	Gavaway : mais je connais....je cherche pas l'emotion mais la sensation c'est tout
Gavaway	cosette_lycee : tué coutes quoi comme musique ?
Mike	ACTION écoute KiKE Dubois - _-Sound of Makina-_

cosette_lycee	un peu de tt saud du ra ! !
cosette_lycee	sauf du rap ! !
Gavaway	ahhh
Mike	re cosette_lycee
Gavaway	tu aimes alain souchon ?
beaugoss	hello
Meublo_Crookah	je t'aime Avrilav, je t tjrs aime
cosette_lycee	re kookoo mike
Avrilav	gohan : ouai ça va
Gilou	c beau l'amour
Puppy	'llo :)
Avrilav	Gavaway : oui
cosette_lycee	tu vois pas comme s'ame pas
cosette_lycee	s'aime
Avrilav	ça va bien toi Meublo_Crookah ?
Gavaway	Mike met toi online sur msn !
Gavaway	yo Whipsy !
Mike	Gavaway : nop, je suis sous linux
Gavaway	Mike : tu fais chier
Meublo_Crookah	kan tu es la oui
Meublo_Crookah	:)
Mike	alors pê quand j'aurais un autre linux moins daubé que red hat
Whipsy	lu
Mike	pourquoi tu veux que je sois sous msn ?
Avrilav	Gavaway : j'apprécie la chanson française tu sais
Mike	laisse tomber le dc
Mike	dcc
Mike	ça passe pas
Gavaway	Mike : pour t'envoyer des fichiers !
Avrilav	Meublo_Crookah : t'as bu, tro fumé ? va te reposer...
Mike	ça prend 1/2 heure pour envoyer un truc par msn pour moi (fw oblige)
Meublo_Crookah	je fume jms trop
Meublo_Crookah	moi meublo
Meublo_Crookah	:)
Anonyme3252753	salut
cosette_lycee	msm on s'en fou !
cosette_lycee	beau gosse t'es brun ? ?
Avrilav	Meublo_Crookah : le vieux délire que tu t'tapes...
Anonyme3252753	salu
beaugoss	oui

cosette_lycee	t'as quel age ?
Anonyme3252753	salut
Meublo_Crookah	kler
Meublo_Crookah	:)
beaugoss	20
cosette_lycee	ouahhh !
cosette_lycee	t ou ?
Mike	ne semble-t-elle pas intéressée ?
Le_Ptit_Chef	le ptit chef le ptif chef un jour viendra, le grand chef le grand chef tu deviendras :p bnap all
cosette_lycee	chur mike ! !
Mike	:)
Meublo_Crookah	cosette_lycee c une chaude du string
cosette_lycee	beaugosse ? ?
Mike	faut encore qu'elle en porte un Meublo_Crookah
cosette_lycee	tu es d'ou stp ! !
Gilou	lol
Avrilav	il est raide..mdrr
Meublo_Crookah	bien vu Mike :)
Mike	si ça se trouve elle n'a rien
Mike	:)
beaugoss	06
cosette_lycee	oh t loin ! ! !
cosette_lycee	merdoum la !
beaugoss	et toi
Meublo_Crookah	oue surement minijupe sans rien
Meublo_Crookah	c normal
Gavaway	cosette_lycee : dis moi , pourquoi tu te cherches pas un copain en vrai ?
cosette_lycee	parceque j'aime bien l'aventure
Mike	Meublo_Crookah : ça permet une plus grande rapidité !
cosette_lycee	c interdit ?
Avrilav	Gavaway : souchon c'est bien beau mais ya bien mieux comme auteur français, toi tu tripes sur les textes en fait ?
Meublo_Crookah	normal, c un acces direct sans file d'attente
Meublo_Crookah	:)
Gavaway	cosette_lycee : c'est pas avec un type à l'autre bout du monde que tu auras des relation sexuelle fréquentes.
Gavaway	Avrilav : exact
Mike	:)
Mike	Meublo_Crookah : ouaip
Gavaway	et pis j'adore ces vieux ttitres
PitiRom	souchon is the leet !
Mike	or not

Avrilav	Gavaway : moi c +renaud,brassens,brel...
Meublo_Crookah	:)
Mike	ACTION écoute Makina - La Conquète De L'ouest
Gavaway	renaud c'est sympa
cosette_lycee	j'cherche ps a l'utre bout du mnde !!
Gavaway	j'aime bien brassens par période
Gavaway	brel aussi
cosette_lycee	j'demande uste un pti rencart
Gavaway	cabrel j'apprécie beaucoup.
beaugoss	souchon il existe vraiment ?
cosette_lycee	oh brel oui !!
PitiRom	lo
PitiRom	llll
Gavaway	cosette_lycee : si tuv eux pas tirer ton coup c'est pas drole.
LightDragon	coucou Whipsy_aie :)
Avrilav	Gavaway : ouai enfin moi tout ça ça me gave vite, je préfére d'autres styles
Meublo_Crookah	cosette_lycee tu me fe une gaterie ?
cosette_lycee	nan mueblo pske je sent ke t pas brun toi
Whipsy_aie	coucou LightDragon
Avrilav	pliéé...

3 Thèmes extraits automatiquement depuis Wikipédia

- Library and information science
  - Reference works
    - Dictionaries
    - Distance education
    - Clients
    - Web sites
    - Encyclopedias
    - Databases
    - Atlases
    - Academic disciplines
    - Prefixes
    - Lists
    - Curricula
    - News agencies
    - Handbooks and manuals
    - Archives
    - Grammar
    - Biographical dictionaries
    - Colleges
    - Culture
    - Reference works in the public domain
    - Knowledge
    - Style guides
    - Universities
    - Research
    - Reading
    - Arts
    - Medical manuals
    - Writing
    - Books
    - Glossaries
    - Government agencies
    - Indices
    - Library cataloging and classification
    - Periodic table
    - Topics
    - Suffixes
    - Information
    - Directories
    - Libraries
    - Trivia books
    - Reference book stubs
    - Search engines
    - Almanacs
- Thought
- Philosophy
  - Aesthetics
  - Ethics
  - Theories
  - Logic
  - Branches
  - Epistemology
  - Literature
  - Philosophers
  - Schools and traditions
  - Movements
  - Arguments
  - History
  - Metaphysics
- Thinking / thinking skills
  - Humanism
  - Ayyavazhi
  - Pantheism
  - Taoism
  - Psychometrics
  - Islam
  - Cao Dai
  - Emotion
  - Ritual
  - Determinism
  - Mysticism
  - Demons
  - Prayer
  - Chinese traditional religion
  - Exorcism
  - New Age
  - Error
  - Belief
  - Perception
  - Confucianism
  - Memory biases
  - Unitarian Universalism
  - Attention
  - Theology
  - Satanism
  - Intelligence
  - Qur'an
  - Theosophy
  - Spiritualism
  - Jainism

Spirituality
Deities
Gnosticism
Animism
Esotericism
Judaism
Imagination
Atheism
Nootropics (smart drugs)
Unitarianism
Christianity
Bible
Mnemonics
Psychological adjustment
Organizational thinking
Problem solving
Creativity
Scientology
Memory
Transcendentalism
Deism
Polytheism
Spiritism
Monotheism
Skepticism
Zoroastrianism
Religious law
Shamanism
Buddhas
Learning
Decision theory
Prophecy
God
Agnosticism
Religious faiths, traditions
Cognition
Atheism
Intelligence researchers
Shinto
Mythology
Buddhism
Bahá'í Faith
Qualities of thought
Rastafarianism
Occult
Wicca
Monism
Cognitive biases
Religion
  Jesus
  Falun Gong
  Paganism
  Mormonism
  Religious ethics
  Tenrikyo
  Hinduism
  Neopaganism
  Sikhism
  Agnosticism
Healthcare occupations
  Nutrition
    Nutritional advice pyramids
    Phytochemicals
    Minerals
    Nutrients
    Vitamins
    Nootropics
    Dietetics
    Dietary supplements
  Medicine
    Veterinary medicine
  Exercise
    Weight training
    Exercise physiology
    Pilates
    Dancing
    Walking
    Bodyweight exercise (Calisthenics)
    Cycling
    Aerobics
    Swimming
    Hiking
    Exercise equipment
    Exercise instructors
    Yoga
    Running
    T'ai Chi Ch'uan
    Sports
    Weight training exercises
  Hygiene
    Cleaning
    Oral hygiene
  Human medicine
    Events
    Timelines
    Neurology
    Human Genetics
    Gastroenterology

- Pathology
- History
- Geriatrics
- Obstetrics
- Gynecology
- Alternative medicine
- Forensics
- Endocrinology
- History of science
- Rheumatology
- Historiography
- Psychiatry
- Urology
- Nephrology
- Surgery
- Ophthalmology
- Oncology
- Sleep
- Gerontology
- Orthopedic surgery
- Hematology
- Pediatrics
- Cardiology

Self-care
- Life extension
- Sexual health
- Health promotion
- Prevention

Positive psychology
- Mental health
- Psychotherapy

Dentistry
- Orthodontics
- Pharmaceuticals policy
- Dental hygiene

Health science
- Pharmacy
- Midwifery
- Public health
- Epidemiology
- Diseases
- Nursing
- Clinical research
- Optometry
- Nutrition

Public health
- Safety
- Healthcare
- Occupational safety and health
- Pharmaceutical industry
- Health law
- Health standards
- Health promotion

Belief
- Mythology
  - Criticism of religion
  - Islamic mythology
  - Social sciences
  - Jewish mythology
  - Buddhist mythology
  - Society
  - Christian mythology
  - Science
  - Abrahamic mythology
  - Hindu mythology

Applied sciences
- Electronics
  - RFID
  - Integrated circuits
  - Telecommunications
  - Avionics
  - Quantum electronics
  - Consumer electronics
  - Embedded systems
  - Signal cables
  - Radio electronics
  - Electronics manufacturing
  - Circuits
  - Semiconductors
  - Electronic design
  - Electrical components
  - Water technology
  - Optoelectronics
  - Connectors
  - Terminology
  - Surveillance
  - Companies
  - Microwave technology
  - Digital media
  - Molecular electronics
  - Digital electronics
- Engineering
  - Systems engineering
  - Structural engineering
  - Bioengineering
  - Electrical engineering
  - Civil engineering
  - Aerospace engineering

```
    Mechanical engineering
    Software engineering
    Materials science
    Environmental engineering
    Chemical engineering
    Nuclear technology
  Computing
    Unsolved problems in computer science
    Artificial intelligence
    Industrial Networks
    Human-computer interaction
    Companies
    Product lifecycle management
    Computing and society
    Languages
    Programming
    Platforms
    Embedded systems
    Data
    Mobile Internet
    Software
    Computer architecture
    Software engineering
    Computer model
    Multimedia
    Free software
    Information systems
    Internet
    Classes of computers
    Real-time computing
    Computer security
    Operating systems
    Computer science
    Networks
  Technology timelines
  Transportation
    Public transport
    Cycling
    Rail transport
    Aviation
    Road transport
    Automobiles
    Water transport
    Spaceflight
    Shipping
    Vehicles
Nature
  Biology
    Neuroscience
    Ecology
    Medicine
    Health sciences
  Physical sciences
    Geology
    Chemistry
    Space
    Physics
    Earth sciences
    Climate
    Astronomy
  Nature
    Natural resources
    Self
    Surnames
    Pollution
    People
    Animals
    Humans
    Plants
    Life
    Environment
    Personal life
  Scientific method
    Scientists
Events
    North America
    Oceania
    Asia
    South America
    Africa
    Europe
    Theorems
    Education
    Numbers
    Geometry
    Proofs
    Measurement
    Analysis
    Equations
    Arithmetic
    Logic
    Abstraction
    Trigonometry
    Mathematics
Abstraction
  Statistics
    Analysis of variance
    Uncertainty of numbers
```

```
    Survival analysis
    Summary statistics
    Data analysis
    Bayesian statistics
    Regression analysis
    Nature
    Experimental design
    Statistical theory
    Categorical data
    Non-parametric statistics
    Multivariate statistics
    Science
    Decision theory
    Logic and statistics
    Parametric statistics
    Natural sciences
    Time series analysis
    Stochastic processes
    Covariance and correlation
    Sampling
Society
  Science
  Social sciences
    Sociology
    Archaeology
    Demographics
    Anthropology
    Information studies
    Psychology
    Media studies
    Economics
    Cultural studies
    Sexology
    Linguistics
    Systems theory
    Political science
    International relations
    Heuristics
    Social scientists
  Society
    Information science
    Information technology
    Internet
    Government
    Business
    Money
    Energy
    Labor
    Military
    Ergonomics
    Biotechnology
    Technology forecasting
    Nanotechnology
    Forestry
    Activism
    Earthquake engineering
    Family
    Space exploration
    Infrastructure
    Communication
    Medicine
    Rights
    Telecommunications
    Sound technology
    Industry
    Industries
    Automation
    Firefighting
    Military science
    Electronics
    Microtechnology
    Nutrition
    Fire prevention
    Home
    Management
    Nuclear technology
    Organizations
    Applied sciences
    Mining
    Marketing
    Design
    Communication
    Chemical engineering
    Blu-ray
    Manufacturing
    Politics
    Forensics
    Health
    Technology
    Optics
    Cartography
    Peace
    Ethnic groups
    Robotics
    Metalworking
    Law
    War
    Finance
```

- Design
- Drawing
- Architecture
- Crafts
- Photography
- Painting

Performing arts
- Theatre
- Opera
- Film
- Storytelling
- Dance
- Music

Arts

Mass media
- Maps
- Continents
- Publications
- Oceans
- Mountains
- Health
- Self-care
- Landforms
- Geography
- Lakes
- Villages
- Newspapers
- Parks
- Rivers
- Countries
- Subterranea
- Radio
- Film
- Publishing
- Healthcare occupations
- Towns
- Television
- Deserts

Humanities
- Philosophy
- Mythology
- Cooking
- Parties
- Celebrities
- Mass media
- Traditions
- Popular culture
- Languages
- Entertainment
- Food and drink
- Toys
- Critical theory
- Games
- Parapsychology
- Tourism
- Radio
- Publications
- Censorship in the arts
- Television
- Literature
- Pets
- Classics

Sports
- American football
- Auto racing
- Skiing
- Ice hockey
- Swimming
- Olympics
- Basketball
- Horse racing
- Gymnastics
- Rugby league
- Canoeing
- Golf
- Whitewater sports
- Cycling
- Lacrosse
- Boxing
- Rugby union
- Cricket
- Sailing
- Soccer
- Baseball
- Tennis

4 Matériel

Pour le développement et les tests, nous utilisons deux configurations différentes.

Serveur. En tant que serveur, nous utilisons une machine dédiée peu puissante :

Processeur	Via C7-D Esther
Architecture	x86, architecture *in-order* (peu adpatée au multi-tâches)
Fréquence	1,5 GHz
Cache L1	128 kio
Cache L2	128 kio
Mémoire vive	1 Gio
BUS	533 MHz
Disque dur	SATA 133 MHz
Système	Linux 2.6 32 bits

Client. En tant que poste de développement et client, nous utilisons une machine plus puissante :

Processeur	Intel Core2 Duo E4600
Architecture	x86-64, architecture *out-of-order* (multi-tâches)
Fréquence	2,4 GHz
Cache L1	64 kio
Cache L2	2 Mio
Mémoire vive	4 Gio
BUS	800 MHz
Disque dur	SATA II 266 MHz
Système	Linux 2.6 64 bits

Table des figures

Liste des tableaux

Résumé

Notre thème de recherche général concerne les aides informatisées au dialogue en langue seconde, oral et/ou écrit. Cette thèse se concentre sur la définition et l'étude, au moyen de corpus et d'un prototype, de procédés d'aide au dialogue écrit (tchat) en langue seconde, dans un contexte de « médiation faible ». Nous présentons dans un premier temps ce qu'est le tchat « ordinaire » en langue première, ses divergences de surface et ses convergences profondes vis à vis des autres formes d'écrit. Nous montrons ensuite les limites des aides actuelles, à « médiation forte », dans laquelle l'outil d'aide est interposé entre des locuteurs supposés totalement incapables de communiquer par un autre biais, de sorte qu'on ne traite jamais le cas pourtant fréquent où ils peuvent utiliser une langue intermédiaire. Nous adaptons au tchat le scénario du projet VerbMobil (1992-2000) et proposons une approche à « médiation faible » pour locuteurs partiellement bilingues, capable de tenir compte de leurs compétences et de leurs problèmes spécifiques.

Le prototype développé dans ce cadre, Koinè, permet d'étudier les contraintes informatiques, ergonomiques et linguistiques d'un tel système, de proposer des solutions, et de les expérimenter. Des aides au dialogue oral ont été prévues, mais, comme la reconnaissance vocale du tout venant au téléphone ou sur IP n'est pas encore assez avancée pour ce type d'utilisation, la version actuelle est centrée sur l'écrit. Koinè est un service Web, construit à l'aide de la bibliothèque logicielle Krater, qui accélère et simplifie le développement d'applications Web. Koinè agrège des fonctionnalités utiles pour surmonter les obstacles de la communication en langue non native, telle que tableau blanc, livre de phrases interactif personnalisable, prétraduction par traduction automatique, mesures d'intelligibilité et de prototypicalité des énoncés, et possibilité de « désambiguïsation interactive et participative ».

Abstract

Our general research topic concerns computer-assistance to chat in a second language, spoken and/or writen. This thesis focuses on the definition and study, through a corpus and a prototype, of informatic assistants for written conversations (chat) in a second language, in a low mediation context. We present first what is "ordinary" chat in first language, the differences of surface and deep similarities towards other forms of writing. Then we show the limitations of current tools, with "strong mediation", in which the tool is interposed between speakers supposedly totally unable to communicate by other means, and that does not consider the case where the speakers may use an intermediary language. We adapt the scenario of Verbmobil project (1992-2000) for chat, and propose an approach with "low mediation" for partially bilingual speakers, able to take into account their skills and their specific problems. With the prototype developed in this view, Koine, we can study the informatic, ergonomic and linguistic constraints of such a system, and propose solutions and experiments. Koine is a Web service, which incorporates useful features to overcome barriers of communication in non-native language, such as whiteboard, interactive customizable phrasebook, MT pretranslation, measures of intelligibility and prototypicality, and possibility of "interactive and participative disambiguation."