

Évaluation du premier démonstrateur de traduction de parole dans le cadre du projet NESPOLE

Solange Rossato, Hervé Blanchon, Laurent Besacier

CLIPS-IMAG
BP 53
38041 Grenoble Cedex 9
{prenom.nom}@imag.fr

Résumé – Abstract

Cet article rapporte les résultats d'un ensemble d'évaluations menées dans le cadre du projet NESPOLE! de traduction de parole. Dans la situation choisie, un client (français, allemand, américain) parle avec un agent de voyage italien (chacun dans sa langue maternelle) pour organiser un séjour en Italie. Cinq séries d'évaluations ont été conduites, toutes sur les mêmes données. La première concerne la reconnaissance automatique de la parole seule. Les deux suivantes concernent la (rétro-) traduction monolingue sur les résultats de la reconnaissance appliquée aux données ainsi que sur la transcription de celles-ci. Les deux dernières concernent la traduction bilingue sur les résultats de la reconnaissance et sur les transcriptions. Le but de cette évaluation était d'analyser les performances de nos systèmes à la fin de la seconde année du projet. Les cinq ensembles de résultats concernant les modules du français sont fournis, commentés et comparés aux résultats obtenus pour les autres langues du projet.

In this paper we present the results of a set of evaluations conducted in the context of the NESPOLE! speech to speech translation project. The chosen situation involves a client (French, German, American) talking to an Italian travel agent (both using their own language) to organize a stay in Italy. Fives series of evaluation were conducted on the same data set. The first series concerned the Automatic Speech Recognition alone. Two other series were about monolingual (back-) translation from ASR outputs on the data set and from transcriptions of the data set. The last ones were about bilingual translation from both the ASR outputs and the transcriptions. The goal of the evaluation was to check the performances of the system at the end of the second year of the project. The fives sets of results concerning the French modules are given, commented and compared to results of the other languages.

Mots Clés – Keywords

Traduction de parole, évaluation
Speech to speech translation, evaluation

Introduction

Le projet NESPOLE!¹, co-financé par l'Union Européenne et la NSF (EU), adresse la problématique de la traduction automatique de parole et ses éventuelles applications dans le domaine du commerce électronique et des services [1]. Les langues impliquées sont l'Italien, le Français, l'Allemand et l'Anglais. Les partenaires sont l'ITC/IRST de Trento (Italie), ISL Labs. de UKA (Karlsruhe, Allemagne) et CMU (Pittsburgh, USA), Aethra (une société italienne spécialisée dans le domaine de la vidéoconférence), APT (une agence de tourisme dans la région du Trentin en Italie) et le laboratoire CLIPS (Grenoble, France).

Le scénario NESPOLE! met en jeu un agent parlant italien, présent dans une agence de tourisme en Italie, et un client qui peut être n'importe où (parlant anglais, français ou allemand) et utilisant un terminal de communication le plus simple possible (PC équipé d'une carte son et d'un logiciel de vidéoconférence type NetMeeting™). Ce choix correspond aux technologies disponibles aujourd'hui, mais, dans un futur proche, les mobiles de troisième génération (UMTS) pourraient éventuellement être utilisés comme terminaux.

Le client veut organiser un voyage dans la région du Trentin en Italie, et navigue sur site Web de APT (l'agence de tourisme) pour obtenir des informations. Si le client veut en savoir plus, sur un sujet particulier, ou préfère avoir un contact plus direct, un service de traduction de parole en ligne lui permet de dialoguer, dans sa propre langue, avec un agent italien de APT. Une connexion, via NetMeeting™, est alors ouverte entre le client et l'agent, et la conversation médiatisée (avec service de traduction de parole) entre les deux personnes peut alors démarrer.

Dans le projet, l'accent est mis sur certains problèmes scientifiques en traduction automatique de parole : robustesse, extensibilité (extension de la couverture d'un domaine) et portabilité (passage d'un domaine à un autre).

Cet article décrit plus particulièrement la campagne d'évaluation des systèmes de traduction de parole, conduite à la fin de la seconde année du projet NESPOLE!.

La *section 1* décrit les modules de traduction développés au CLIPS, pour le français. La méthodologie d'évaluation est décrite dans la *section 2* de cet article. À notre connaissance, très peu d'articles ont déjà abordé le problème de l'évaluation en traduction de parole. La *section 3* décrit les résultats de l'évaluation conduite à la fin de la première période du projet NESPOLE!. Finalement, conclusions et perspectives sont données à la fin de cet article.

1 Modules de traduction de parole

1.1 Approche par Langage pivot (IF)

L'IF (Interchange Format) [2,3] est fondé sur des actes de dialogue (DA²) auxquels sont adjoints des arguments. Un acte de dialogue est constitué d'un acte de parole (SA³) complété de concepts. Les actes de dialogue décrivent les intentions, les besoins de celui qui parle (*give-information, introduce-self, ...*). Les concepts précisent à propos de quoi l'acte de dialogue est exprimé (*price, room, activity, ...*). Les arguments permettent d'instancier les valeurs des variables du discours (*room-spec, time, price, ...*).

¹ voir <http://nespole.itc.it/>

² Dialogue Act

³ Speech Act

Pour un client dont un tour de parole signifie "je voudrais la chambre à 70 euros" l'IF est :

```
c: give-information+disposition+price+room
( disposition=(who=i, desire),
  price=(quantity=70, currency=euro),
  room-spec=(identifiability=yes, room)
)
```

c: indique ce c'est le client qui parle ; give-information+disposition+price+room est l'acte de dialogue ; disposition, price, room-spec sont des arguments supérieurs (top level arguments) ; who, quantity, currency, identifiability sont des arguments inférieurs (embedded arguments) ; i, desire, 70, euro, yes, room sont des valeurs.

L'architecture d'un système de traduction utilisant l'approche pivot est décrite dans la *figure 1*.

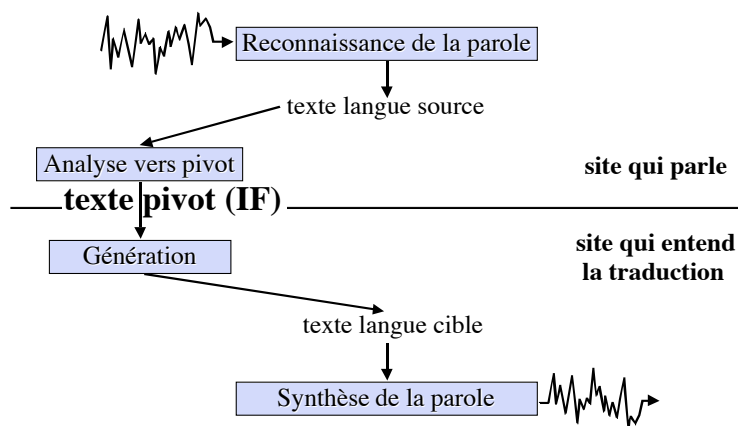


Figure 1 Interaction entre les modules de traduction de parole dans l'architecture IF.

L'avantage le plus évident de l'approche par pivot (IF) est la réduction du nombre de modules à réaliser. Si n langues sont impliquées, deux modules, d'analyse et de génération vers/depuis l'IF, pour chaque langue permettent la traduction pour toutes les paires de langues possibles. Si une nouvelle langue vient s'ajouter au projet, il suffit alors de développer les modules d'analyse et de génération vers/depuis l'IF pour cette langue afin qu'elle puisse être intégrée avec les n autres.

Le défaut de cette approche réside dans la difficulté à définir le langage pivot, les concepts qui sont couverts, ainsi que sa syntaxe. Cela est vrai même lorsque le domaine est limité à une tâche particulière comme l'information touristique.

1.2 Reconnaissance automatique de la parole

La première étape d'un système de traduction de parole est la reconnaissance automatique de la parole. Notre système de reconnaissance RAPHAEL utilise la boîte à outils Janus-III [3] de CMU. Les modèles acoustiques dépendants du contexte sont appris sur un corpus contenant 12 heures de parole continue de 72 locuteurs extrait de la base de données Bref80 [4].

Le vocabulaire est d'environ 20000 mots ; une partie est spécifique à la réservation d'hôtels et à l'information touristique, et le reste du vocabulaire correspond aux mots les plus fréquents du français (d'après un comptage de mots fait sur le Web [5]).

Le modèle statistique de langage est un modèle de trigrammes calculé sur un large corpus extrait de documents textuels collectés automatiquement sur le Web. RAPHAEL est décrit plus en détail dans [6] et [7].

1.3 Analyse Français vers IF

Le module d'analyse actuellement utilisé [8] met en œuvre une méthode, fondée sur des patrons, similaire à ce qui est décrit dans [9]. Cette approche est bien adaptée à l'analyse d'une entrée bruitée ainsi qu'à la fabrication des constituants de l'IF. Les patrons permettent d'exploiter des segments du texte produit par le module de reconnaissance en autorisant des insertions, des effacements et des accords en genre et en nombre illégaux (exemple d'entrée "je veux quatre quatre chambre heu double").

L'analyse s'effectue en quatre étapes.

- Un tour de parole est d'abord découpé en Unités Sémantiques de Dialogue (SDU⁴). Les SDUs sont ensuite traitées en séquence. Pour trouver des bornes de SDU nous utilisons des groupes simples (de confirmation, d'acquiescement, de refus, d'excuse, ...) et des articulations du discours (conjonctions suivies d'un pronom personnel ou d'un verbe).
- Pour chaque SDU, un domaine est calculé (*activity*, *accommodation*, *attraction*, ...). Ces domaines servent à délimiter le nombre et le type des arguments que l'on va chercher à repérer.
- Selon le domaine, les arguments autorisés présents sont instanciés. Chaque argument est décrit par un ensemble de groupes syntagmatiques modèles qui utilisent des variables de classes (noms d'hôtels, types d'hôtels, types de chambres, ...).
- Finalement, l'acte de dialogue est calculé en prenant en compte les arguments instanciés ainsi que d'autres marqueurs présents dans l'entrée (attitudes, question, négation, ...).

Des patrons sont utilisés à plusieurs stades du processus, et principalement pour décrire le découpage en SDUs ainsi que les réalisations possibles des arguments.

1.4 Génération IF vers Français

Pour la génération, nous avons choisi de mettre en œuvre une approche par règles avec Ariane-G5 [10]. Pour ce faire, nous utilisons les fichiers de spécifications de l'IF afin de produire automatiquement des squelettes de grammaires et de dictionnaires. Une IF en entrée de ce module de génération est transformée en un arbre linguistique qui est transmis à un module de génération à usage général.

Lorsque ce module de génération sera prêt, nous aurons une couverture complète de l'IF. Les inconvénients de cette approche sont d'une part, la lourdeur du processus d'initialisation (l'IF décrit beaucoup d'événements potentiels), et d'autre part, les changements réguliers faits à la spécification.

Nous développons en parallèle un module de génération fondé sur des modèles de phrases. Cette approche, plus "rustique", se justifie par la structure de l'IF et sa faible profondeur linguistique.

Dans cette approche, des "phrases à trous" modèles sont associées à des familles d'actes de dialogue. Les trous sont remplis avec les syntagmes français qui correspondent aux valeurs des arguments présents dans l'IF.

⁴ Semantic Dialogue Unit, unité maximum du texte à analyser qui puisse être représentée par une IF.

1.5 Synthèse de la parole

La dernière étape en traduction de parole est la synthèse de parole. Le système Euler⁵ de l'université Polytechnique de Mons est utilisé. La qualité de la synthèse d'Euler est suffisante pour notre application. Cependant, l'évaluation du module de synthèse n'est pas considérée dans cet article.

2 Méthodologie d'évaluation

La reconnaissance de la parole a été évaluée de manière indépendante. Nous avons ensuite évalué différentes combinaisons des modules d'analyse et de génération, considérées comme des boîtes noires, sur deux classes de données. La première classe de données est constituée de transcriptions des tours de paroles utilisés (appelées ensuite **références**). La seconde classe est constituée des sorties du module de reconnaissance automatique de la parole sur les tours de parole considérés.

2.1 Évaluation séparée des modules

Tout d'abord, une évaluation séparée des composants a été conduite. Les procédures d'évaluation pour chaque composant sont décrites dans les paragraphes suivants.

2.1.1 Évaluation de la reconnaissance automatique de la parole

Nous utilisons bien sûr, la mesure standard du taux d'erreur (ou word-error-rate WER).

Cependant, ce critère ne prend pas en compte le fait que certaines erreurs de reconnaissance peuvent avoir des conséquences plus ou moins importantes sur la qualité du système complet de traduction de parole. Ainsi, nous avons aussi mis en œuvre une évaluation objective de la qualité des sorties du module de reconnaissance. La sortie du système de reconnaissance est évaluée comme si elle représentait une paraphrase de la transcription d'origine (manuelle) du signal.

Cela est fait en utilisant les mêmes critères d'évaluation que ceux qui sont utilisés pour évaluer la qualité de traduction (décrits plus loin dans la *section 2.3*).

2.1.2 Évaluation de la traduction monolingue sur les références

Les modules d'analyse de la langue X vers l'IF et de génération de l'IF vers la langue X sont ici considérés comme un module unique de traduction X vers X monolingue qu'il faut évaluer. L'évaluation de la traduction monolingue est faite de la façon suivante□analyse de la phrase d'origine en langue X vers l'IF, suivie de la retro-génération en langue X à partir de l'IF. La sortie produite est alors comparée à la phrase d'origine par des évaluateurs humains. Cette évaluation peut être faite par n'importe quel locuteur natif de la langue X.

2.1.3 Évaluation de la traduction bilingue sur les références

Les modules d'analyse de la langue X vers l'IF et de génération de l'IF vers la langue Y sont ici considérés comme un module unique de traduction X vers Y qu'il faut évaluer. L'évaluation de la traduction est faite de la façon suivante□analyse de la phrase d'origine en langue X vers l'IF, suivie de la génération en langue Y à partir de l'IF. La sortie produite est alors comparée à la phrase d'origine par des évaluateurs humains. Cette évaluation ne peut être faite que par des locuteurs bilingues.

⁵ <http://tcts.fpms.ac.be/synthesis/euler/>

2.2 Évaluation de la chaîne de traduction de parole

2.2.1 Évaluation de la traduction monolingue sur les sorties de la reconnaissance automatique de la parole

Pour l'évaluation monolingue complète, le système de reconnaissance automatique de la parole de la langue X est combiné avec les modules de d'analyse et de génération depuis et vers cette langue X. L'entrée est un signal dans la langue X qui est reconnu par le système de reconnaissance. L'hypothèse de reconnaissance est alors analysée en IF puis retro-générée en langue X. Le chaîne textuelle ainsi obtenue est alors comparée à la phrase d'origine (transcription manuelle du signal de départ).

Ceci permet une évaluation complète de toute la chaîne de traitement pour le langage X. Chaque évaluation est faite indépendamment par chaque partenaire du projet.

2.2.2 Évaluation de la traduction bilingue sur les sorties de reconnaissance automatique de la parole

Des évaluations complètes ont été aussi conduites entre les différents partenaires. Toutes les phrases de l'évaluation sont d'abord passées dans les modules de reconnaissance et d'analyse pour une langue X donnée, produisant ainsi un corpus de chaînes IF. Ce corpus est ensuite envoyé au second site de la langue Y qui applique son module de génération pour obtenir un corpus de phrases en langue Y.

Des évaluateurs humains bilingues réalisent alors l'évaluation en comparant les transcriptions correctes (manuelles) des signaux en langue X avec les phrases traduites en langue Y.

2.3 Procédure d'évaluation

Trois évaluateurs ont participé à chaque évaluation. L'évaluation est faite, non pas au niveau de chaque tour de parole ou phrase, mais au niveau des SDUs. Ainsi, avant l'évaluation, les données à traduire sont d'abord segmentées manuellement en SDU. Pour chaque SDU et sa version traduite, l'évaluateur doit juger de la qualité de traduction en donnant une note. Trois choix sont proposés pour la note.

- **p** pour *perfect* si l'évaluateur estime que la qualité de traduction est très bonne
- **k** pour *okay* si l'évaluateur estime que la qualité de traduction est acceptable
- **b** pour *bad* si l'évaluateur estime que la qualité de traduction n'est pas acceptable

3 Résultats d'évaluation

3.1 Données d'évaluation

Quatre dialogues extraits de la base de données NESPOLE! [11] ont été utilisés. Deux d'entre eux correspondent à une conversation entre un client qui veut organiser des vacances d'hiver dans le Val-di-Fiemme en Italie et un agent italien du bureau de tourisme APT. Les deux autres conversations ont la même configuration client/agent, mais concernent l'organisation de vacances d'été dans la même région.

Les signaux de parole en français ont été re-enregistrés à partir des transcriptions des tours de parole du client de ces quatre dialogues (fréquence d'échantillonnage de 8kHz et codage G711 compatibles avec le format de vidéoconférence). Ces données représentent 235 signaux

correspondant aux 235 tours de parole du client dans les quatre dialogues, prononcés par deux locuteurs différents (1 homme, 1 femme).

Finalement, les transcriptions de ces 235 tours de parole sont segmentées manuellement en 427 unités sémantiques (SDUs) pour l'évaluation de la traduction. Après avoir appliqué les modules de reconnaissance et/ou de traduction sur ces données, des évaluateurs humains jugent de la qualité des SDUs traduites et les comparant aux SDUs d'origine, comme cela est décrit dans la section précédente. Dans tous les tableaux des paragraphes 3.2 à 3.6, les pourcentages de chaque note sont donnés pour chaque évaluateur et pour chaque dialogue. Dans chaque case, le premier chiffre représente le pourcentage de SDUs jugés comme correctement traduites (p+k cumulées) et le second chiffre, entre parenthèses, représente le pourcentage de SDUs jugées comme parfaitement traduites (p seulement).

Un vote majoritaire est aussi appliqué pour chaque SDU jugée par les trois évaluateurs. Dans ce cas, une SDU n'est gardée que si au moins deux évaluateurs parmi trois donnent la même note. Ainsi, les SDU qui donnent lieu à des différences de jugement entre les évaluateurs sont supprimées de l'évaluation. Les résultats de ce vote majoritaire sont donnés sur la dernière ligne de chaque tableau.

3.2 Résultats d'évaluation du système de reconnaissance

Le taux d'erreurs de reconnaissance (WER) obtenu sur les 235 signaux correspondant à 235 tours de parole de clients, est de 28.8% (soit 71.2% de mots correctement reconnus). Nous rappelons que le vocabulaire comporte 20000 mots (c.f. § 1.2).

Les résultats de l'évaluation humaine des sorties du système de reconnaissance en tant que traduction par paraphrase de la phrase d'origine (référence) sont donnés dans la *Table 1*.

Environ 65% des SDUs sont jugées correctes par les évaluateurs humains. Cette seconde évaluation est plus informative pour le système de traduction que le taux d'erreur conventionnel car 65% de SDUs correctes signifie que dès cette phase de reconnaissance, nous sommes sûrs que 35% des données ne pourront pas être correctement traduites. Cette phase d'évaluation est également un bon moyen de juger de la qualité des évaluateurs et de vérifier qu'ils rendent des jugements cohérents sur la majorité des données (ce qui est le cas ici).

3.3 Résultats d'évaluation de la traduction monolingue sur les références

L'évaluation de la traduction (analyse + génération) est réalisée par des évaluateurs humains. Les résultats d'évaluation se retrouvent dans la *Table 2*.

Environ 55% des SDUs traduites sont jugées correctes. Les résultats du vote majoritaire montrent une bonne cohérence entre les évaluateurs. Après cette phase d'évaluation, nous sommes sûrs que 45% des SDUs ne pourront être correctement traduites par le système complet de traduction de parole. Il est alors important de savoir si ces SDUs mal traduites contiennent les SDUs mal reconnues par le système de reconnaissance, ou si les erreurs se cumulent lorsqu'on cascade les deux systèmes. C'est ce qui est présenté dans la section suivante.

3.4 Résultats d'évaluation de la traduction de parole monolingue sur les sorties de la reconnaissance automatique de la parole

Les résultats sont présentés dans la *table 3*. Nous voyons ici, que la chaîne complète de traduction de parole permet de traduire correctement environ 41% des SDUs. Ce résultat, seul, serait difficile à interpréter, mais les évaluations décrites dans les *sections 3.2 et 3.3* montrent la contribution respective des modules de reconnaissance et de traduction séparés. C'est une

information importante qui permet de savoir d'où viennent les erreurs et d'améliorer en conséquence le ou les modules concernés.

3.5 Résultats d'évaluation de la traduction bilingue sur les références

Les résultats d'évaluation de la traduction français vers italien, sont présentés dans la *table 4*. Il semblerait que pour l'évaluation bilingue, le comportement des évaluateurs n'est pas aussi cohérent que pour la traduction monolingue. Ceci peut être dû à différents niveaux d'expertise des évaluateurs dans une des deux langues.

De plus, le nombre moyen de SDUs jugées correctement traduites est plus bas que pour la même évaluation en mode monolingue (44% en bilingue contre 55% en monolingue). Ceci est dû au fait que les modules du français d'analyse et de génération sont développés conjointement.

Ainsi, le module de génération produira plus facilement du français à partir d'IFs issues de l'analyseur français, tandis que le générateur d'italien sera moins optimal pour produire du texte italien à partir de ces mêmes IFs.

On voit bien là que le problème de la spécification du langage pivot, l'IF, est très important. Quelques incohérences restant dans la spécification de l'IF entre les partenaires sont donc également une explication de cette baisse de résultats en mode bilingue.

3.6 Résultats d'évaluation de la traduction de parole bilingue sur les sorties de la reconnaissance automatique de la parole

Les résultats d'évaluation de la chaîne complète de traduction de parole français vers italien, sont présentés dans la *table 5*. Là encore, le nombre moyen de SDUs jugées comme correctement traduites est légèrement plus faible que pour la même évaluation en mode monolingue. Nous pouvons voir quand même qu'environ 1/3 des données de parole sont correctement traduites du français vers l'italien.

3.7 Cohérence des scores entre évaluateurs

Le vote majoritaire permet d'observer la cohérence des scores des différents évaluateurs. En effet, une SDU est écartée si les trois évaluateurs ont choisi trois scores différents. Ainsi les SDUs conservées sont celles pour lesquelles un consensus peut être obtenu par majorité.

Tâche	% de SDUs conservées avec le vote majoritaire
Reconnaissance (<i>Table 1</i>)	97,2
Traduction monolingue sur références (<i>Table 2</i>)	92,5
Traduction monolingue sur reconnaissance (<i>Table 3</i>)	89,5
Traduction bilingue sur références (<i>Table 4</i>)	86,7
Traduction bilingue sur la reconnaissance (<i>Table 5</i>)	92,0

Pour la reconnaissance, presque toutes les SDU ont été conservées (13 ont été écartées). Pour la traduction, on a une cohérence autour de 90%. Étant donné la taille des échantillons, et surtout le petit nombre d'évaluateurs, nous ne pouvons pas dire si ces chiffres sont statistiquement significatifs. Ils ne sont qu'indicatifs.

3.8 Comparaison avec les autres langues du projet

Nous présentons, dans la *table 6*, un résumé des résultats d'évaluation monolingue et bilingue sans vote majoritaire⁶ (traduction de texte seule et traduction de parole complète) pour toutes les autres langues du projet. Nous pouvons voir que les résultats obtenus pour le français sont tout à fait dans la moyenne des résultats obtenus pour les autres langues. Une description plus détaillée des systèmes des différents partenaires est donnée dans [12].

Conclusion

Dans cet article, nous avons présenté une méthodologie et des résultats complets d'évaluation pour la traduction automatique de parole. À notre connaissance, c'est un des premiers articles qui adresse le problème d'évaluation dans un domaine nouveau comme la traduction automatique de parole.

En effet, jusqu'à maintenant, seules des démonstrations de traduction présentées au public faisaient office de validation de tels systèmes. Avec une procédure d'évaluation bien établie, nous entrons dans une nouvelle phase plus mature du domaine. Nous mesurons ainsi qu'il reste encore beaucoup de problèmes à résoudre et de chemin à parcourir pour avoir des systèmes de traduction utilisables dans des services réels, puisque seulement 30 à 40% des données sont correctement traduites, quelle que soit la langue et le partenaire de NESPOLE! concerné.

Remerciements

Les auteurs remercient les évaluateurs bilingues et monolingues (A.C. Descalle, F. Tajariol, D. Vaufreydaz, R. Lamy, S. Mazenot) et l'équipe de traduction du laboratoire IRST en Italie (qui a généré l'italien à partir de nos IFs), pour leur contribution à cette article.

Références

- [1] Lazzari G. (2000) Spoken translation: challenges and opportunities. Proc. ICSLP'2000. Beijing, China. Oct. 16-20, 2000, vol. 4/4 : pp. 430-435
- [2] Levin L. & al. (1998) *An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues*. Proc. ICSLP'98. Sydney, Australia, 30th November – 4th December, 1998, vol. 4/7: pp. 1155-1158.
- [3] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. (1993) *Recent Advances in JANUS: A Speech Translation System*. Proc. Eurospeech 93. Berlin, Germany, vol. 2: pp. 1295-1298.
- [4] Lamel, L.F., Gauvain, J.L., Eskénazi, M. (1991) *BREF, a Large Vocabulary Spoken Corpus for French*. Proc. Eurospeech 91. Gênes, Italy, 24-26 September, 1991, vol. 2: pp. 505-508.
- [5] Besacier L., Blanchon H., Fouquet Y., Guilbaud J.P., Helme S., Mazenot S., Moraru D., Vaufreydaz D. (2001) *Speech Translation for French in the NESPOLE! European Project*. Proc. Eurospeech 01. Aalborg, Denmark, September 3-7, 2001, vol. 2/4: pp. 1291-1294.

⁶ Les chiffres du vote majoritaires ne sont pas disponibles pour l'instant.

- [6] Vaufreydaz D., Besacier L., Bergamini C., Lamy R. (2001) *From generic to task-oriented speech recognition: French experience in the NESPOLE! European project*, presented at ITRW Workshop on Adaptation Methods for Speech Recognition. Sophia-Antipolis, France 29-30 August, 2001.
- [7] Vaufreydaz D., Besacier L., Bergamini C., Lamy R. (2001) *From generic to task-oriented speech recognition: French experience in the NESPOLE! European project*. Proc. ITRW Workshop on Adaptation Methods for Speech Recognition. Sophia-Antipolis, France. August 29-30, 2001.
- [8] Blanchon H. (2002) *A Pattern-Based Analyzer for French in the Context of Spoken Language Translation: First Prototype and Evaluation*. Proc. COLING'02. Taipei, Taiwan, August 24-September 1, 2002, *Submitted*.
- [9] Zong, C., Huang, T. and Xu, B., (2000). *An Improved Template-Based Approach to Spoken Language Translation*. Proc. ICSLP 2000. Beijing, China. Oct. 16-20, 2000. vol. 4/4: pp.440-443.
- [10] Boitet C. (1997) *GETA's methodology and its current development towards networking communication and speech translation in the context of the UNL and C-STAR projets*. Proc. PACLING-97. Ome, Japan. 2-5 September, 1997. vol. 1/1: pp. 23-57.
- [11] Burger, S., Besacier, L. Metze, F., Morel, C., Coletti, P. (2001) *The NESPOLE! VoIP dialog database*, Eurospeech 2001. Aalborg, Danemark, September 3-7, 2001
- [12] Lavie A., Metze F., Pianesi F., Burger S., Gates D., Levin L., Langley C., Peterson K., Schultz T., Waibel A., Wallace D., McDonough J., Soltau H., Laskowski K., Cattoni R., Lazzari G., Mana N., Pianta E., Costantini E., Besacier L., Blanchon H., Vaufreydaz D., Taddei L. (2002) *Enhancing the Usability and Performance of Nespole! – a Real-World Speech-to-Speech Translation System*. Proc. HLT 2002. San Diego, California. 6p.

Tables de résultats

Table 1 Résultats d'évaluation du système de reconnaissance

%acceptable (%parfait)	Dial. A1 60 ph., 109 SDUs	Dial. A2 74 ph., 139 SDUs	Dial. C3 64 ph., 101 SDUs	Dial. C4 37 ph., 78 SDUs	Tous les dialogues 235 ph., 427 SDUs
Evaluateur 1 (lb)	74.5 (70.0)	49.3 (45.0)	68.0 (61.2)	81.0 (78.5)	66.0 (61.3)
Evaluateur 2 (sr)	70.9 (68.2)	46.8 (40.4)	64.1 (59.2)	80.0 (78.8)	63.1 (59.0)
Evaluateur 3 (rl)	73.6 (70.0)	47.1 (42.1)	69.9 (60.2)	82.5 (77.5)	65.8 (60.0)
Moyenne	73 (69.4)	47.7 (42.5)	67.3 (60.2)	81.2 (78.3)	65.0 (60.1)
Vote_maj	72.5 (69.7)	46.0 (41.0)	68.3 (61.4)	76.9 (75.6)	63.7 (59.5) [415SDUs]

Table 2 Résultats d'évaluation de la traduction monolingue sur les références

%acceptable (%parfait)	Dial. A1 60 ph., 109 SDUs	Dial. A2 74 ph., 139 SDUs	Dial. C3 64 ph., 101 SDUs	Dial. C4 37 ph., 78 SDUs	Tous les dialogues 235 ph., 427 SDUs
Evaluateur 1 (lb)	67.7 (48.6)	44.7 (34.0)	49.5 (36.9)	64.6 (49.4)	55.3 (41.2)
Evaluateur 2 (sr)	63.6 (50.9)	44.7 (38.3)	45.6 (40.8)	56.4 (52.6)	51.8 (44.7)
Evaluateur 3 (rl)	67.6 (55.9)	47.9 (36.6)	49.5 (41.7)	62.0 (51.9)	55.9 (45.5)
Moyenne	66.3 (51.8)	45.8 (36.3)	48.2 (39.8)	61.0 (51.3)	54.3 (43.8)
Vote_maj	61.5 (52.3)	40.3 (36.0)	48.5 (40.6)	53.8 (47.4)	50.7 (43.3) [395SDUs]

Table 3 Résultats d'évaluation de la traduction de parole monolingue sur les sorties de reconnaissance de la parole

%acceptable (%parfait)	Dial. A1 60 ph., 109 SDUs	Dial. A2 74 ph., 139 SDUs	Dial. C3 64 ph., 101 SDUs	Dial. C4 37 ph., 78 SDUs	Tous les dialogues 235 ph., 427 SDUs
Evaluateur 1 (lb)	49.1 (31.3)	26.4 (15.0)	37.3 (24.5)	52.5 (32.5)	39.6 (24.7)
Evaluateur 2 (sr)	54.6 (33.6)	28.2 (18.3)	40.8 (29.1)	57.7 (42.3)	43.2 (29.1)
Evaluateur 3 (rl)	51.8 (34.5)	27.0 (19.9)	35.9 (30.1)	51.3 (46.3)	39.9 (30.9)
Moyenne	51.8 (33.1)	27.2 (17.7)	38.0 (27.9)	53.8 (40.4)	40.9 (28.2)
Vote_maj	49.5 (31.2)	24.5 (14.4)	36.6 (27.7)	46.2 (37.2)	37.7 (26.0) [382SDUs]

Table 4 Résultats d'évaluation de la traduction bilingue sur les références français vers italien

%acceptable (%parfait)	Dial. A1 60 ph., 109 SDUs	Dial. A2 74 ph., 139 SDUs	Dial. C3 64 ph., 101 SDUs	Dial. C4 37 ph., 78 SDUs	Tous les dialogues 235 ph., 427 SDUs
Evaluateur 1 (an)	56.0 (40.4)	34.8 (24.1)	40.8 (31.1)	50.0 (36.3)	44.3 (32.1)
Evaluateur 2 (fe)	51.4 (38.5)	25.0 (21.3)	34.3 (30.4)	50.7 (42.5)	38.6 (31.7)
Evaluateur 3 (sy)	59.8 (47.7)	37.9 (30.7)	43.1 (37.3)	56.3 (46.3)	48.0 (39.4)
Moyenne	55.7 (42.2)	32.6 (25.4)	39.4 (32.9)	52.3 (41.7)	43.6 (34.4)
Vote maj	48.6 (37.6)	27.3 (23.7)	39.6 (34.7)	41.0 (34.6)	38.2 (31.9) [370SDUs]

Table 5 Résultats d'évaluation de la traduction de parole bilingue sur les sorties de reconnaissance de la parole français vers italien

%acceptable (%parfait)	Dial. A1 60 ph., 109 SDUs	Dial. A2 74 ph., 139 SDUs	Dial. C3 64 ph., 101 SDUs	Dial. C4 37 ph., 78 SDUs	Tous les dialogues 235 ph., 427 SDUs
Evaluateur 1 (an)	39.1 (27.3)	23.4 (15.6)	32.0 (24.3)	46.3 (36.3)	33.6 (24.4)
Evaluateur 2 (fe)	43.5 (35.2)	20.1 (11.5)	30.7 (23.8)	46.8 (36.7)	33.5 (25.1)
Evaluateur 3 (sy)	39.1 (29.1)	22.9 (19.3)	33.0 (27.2)	51.3 (41.3)	34.6 (27.7)
Moyenne	40.6 (30.5)	22.1 (15.5)	31.9 (25.1)	48.1 (38.1)	33.9 (25.7)
Vote_maj	37.6 (28.4)	18.7 (15.1)	30.7 (25.7)	42.3 (35.9)	30.7 (24.8) [393SDUs]

Table 6 Comparaison avec les autres langues dans Nespole

Monolingue (%acceptable) Trad / Rec+Trad	Anglais-Anglais 58% / 45%	Allemand-Allemand 31% / 25%	Français-Français 54% / 41%	Italien-Italien 61% / 48%
Bilingue (%acceptable) Trad / Rec+Trad	Anglais=>Italien 55% / 43%	Allemand=>Italien 32% / 27%	Français=>Italien 44% / 34%	
	Italien=>Anglais 47% / 37%	Italien=>Allemand 47% / 31%	Italien=>Français 40% / 27%	