NESPOLE!'s Multilingual and Multimodal Corpus

Erica Costantini[1], Susanne Burger[2], Fabio Pianesi[3]

[1]Department of Psychology, University of Trieste, Italy
costanti@psico.units.it
[2]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, USA
sburger@cs.cmu.edu
[3]ITC-irst, Trento, Italy
pianesi@irst.itc.it

Abstract

NESPOLE! is a EU/NSF jointly funded project exploring multilingual (speech-to-speech translation) and multimodal communication in e-services. The current system allows users speaking different languages (English, French, German and Italian) to interact on the tourism domain through the Internet using thin terminals (PCs with sound and video cards and H323 video-conferencing software). Web pages and maps can be shared among users, by means of a special White Board. NESPOLE! provides for multimodal communication by allowing users to perform gestures on displayed maps, by means of a tablet and a pen. To test the integration of multilinguality with multimodality, and the impact of the latter on the former, we designed and executed an experiment, involving 35 subjects, 28 playing the role of customers (English and German) and 7 playing the role of agents (Italian). Subjects communicated through the NESPOLE! system to accomplish an assigned task (booking an hotel), meeting specific constraints as to available budget, location, distance from relevant spots, etc. Two experimental conditions were considered and compared, differing as to whether multimodal resources were available: a speech-only condition (SO), and a multimodal condition (MM). This paper reports on the resulting corpus, and on the results of the experiment.

1. Introduction

NESPOLE! (NEgotiating through SPOken Language in E-commerce) is a jointly EU/NSF funded project exploring speech-to-speech translation (STST) in e-commerce and e-service sectors (Lazzari, 2000; Lavie et al. 2002; Metze et al. 2002)). The languages addressed in this project are Italian, German, English and French. The scenario for the first showcase, now released, involves an Italian-speaking agent located in an Italian tourism agency (APT), and an English-, German- or French-speaking customer located at an arbitrary location. The two communicate through the Internet using thin terminals (PCs with sound and video cards and H323 video-conferencing software). NESPOLE! provides for multimodal communication too, allowing users to perform gestures on displayed maps, by means of a tablet and a pen.

During an experiment involving three of the partners (Italy, Germany and USA), a multilingual and multimodal corpus was produced in the context of a 'true' speech-to-speech translation scenario. The paper focuses on this corpus and the results of the experiment.

2. Experimental Hypothesis and Design

Previous research using WoZ technique identified performance advantages when interacting with maps multimodally rather than unimodally – including faster task completion, fewer input disfluences, briefer and less complex language, greater satisfaction (Oviatt, 1997a). Moreover, it has been found that multimodal interaction occurs more frequently in case of spatial location commands (Oviatt et al. 1997b). In addition, some studies suggest that well designed multimodal systems can integrate complementary modalities in a manner that supports significant levels of mutual disambiguation of errors (Oviatt. 1999).

It is not clear to which extent these findings can be replicated when 'real' systems for multilingual human-to-human communication are at stake. Real systems, in fact, can introduce disturbing factors such as system's failures, time-lag due to network traffic, etc., which can dilute, weaken, and even contrast the positive effects of multimodality. We therefore designed and executed an experiment, which, on the basis of those researches, aimed to test:

- whether multimodality increases the probability of successful interaction, even with prototypes of 'real' multilingual systems, when spatial information is the focus of the communicative exchange;
- whether multimodality helps decreasing ambiguities and disfluences;
- whether it supports a faster recovery from recognition and translation errors.

To these ends, we devised two experimental conditions:

- a speech-only condition (SO), involving multilingual communication and the possibility for users to share images;
- a multi-modal condition (MM), where users could additionally perform pen-based gestures (pointing, area selection, connection between different areas) on shared maps to convey spatial information.

3. Scenario and Experimental Setting

The scenario of the experiment was modeled after one of the five different tourism scenarios studied during Nespole!'s training data collection (Burger, 2001), enriched with spatial information. It features a customer browsing the web pages of a tourist office in Trentino, Italy. When the customer wants more information, she clicks on a special button, which opens a direct, STST-mediated connection with a human agent. The customer's task was to choose an appropriate location and a hotel within constraints specified a priori concerning the

relevant geographical area, the available budget, etc. The agent's task was to provide the necessary information.

Two kinds of participants were involved: American English native speakers (located at CMU, Pittsburgh) and German native speakers (located at UKA, Karlsruhe), who played the role of the customers, and Italian native speakers (located at Irst, Trento), who were trained to act as tourist agents.

During the experiment, subjects wore a head-mounted microphone, using it in a push-to-talk mode. In the MM condition they drew gestures on maps by means of a tablet-pen device. Each subject could only hear the translated message of the other party (original audio was disabled).

NESPOLE!'s screen displayed three windows:

1. The Aethra White Board window, set at 600x600, which was used to display maps, both for MM and SO condition.
2. The Feedback Window, which displays useful information for the users concerning: the hypothesis string produced by the speech recogniser, and a string informing about the system understanding.
3. The NetMeeting® window, allowing control over the usual features of this application. It has a button to activate/de-activate the microphone (push-to-talk).

A research assistant assisted the participants during the experimental session. Customers received written information and instructions about the scenario, the task, system functionalities and interaction modalities (task and instructions are available on the Nespole! website). Before starting the interaction, we asked customers to write down the information they thought they would need to ask the agents for in order to help customers planning the conversation. In the MM condition, we demonstrated them the White Board functionalities, and allowed them few minutes to familiarize with the optical pen.

Agents were trained by Irst and instructed about how they would better answer (kinds of answers allowed, style, so as to adhere as much as possible to what 'real' agents usually do). Agents' training took longer than customers', since the former had to be more acquainted with the functionalities of the White Board (in a real setting, it can't be required that customers be experienced with the White Board and the pointing devices, whereas this should be part of the agent's skills), and be proficient in the task of searching and providing the requested information. Agents were given description cards with information about two resorts in Val di Fiemme (a tourist resort in Trentino), and three hotels for each place. Only agents were allowed to send maps and webpages to support the interaction. Again, this replicates a 'real' setting, where it is the agent who knows what map or figure can be helpful at a certain point, where to find it, etc.

4. Multimodality

Multimodality in Nespole! is accomplished by the integration of speech and pen-based gestures. The users were allowed to draw gestures on maps loaded on the White Board only during the MM condition. The White Board drawing functionalities include:

- free-hand strokes: the user can draw arrows, lines, circles and other free-hand stokes of her choice;
- lines: the user can connect two point on the maps through a (possibly arrow-headed) line;
- selection of areas on map: this can be done by enclosing portions of maps in elliptical/rectangular figures.

The drawings are performed by means of the pointing device. In addition, appropriate colors can be selected among the palette for all types of drawings, to distinguish among different gestures.

5. Recordings, Transcriptions and Annotations

We scheduled 53 experimental sessions. Eventually, only 47 did actually take place, 19 of which were canceled because of technical problems (system crashes, network failures, etc.) or incomplete recordings. The resulting dialogue corpus therefore consists of 28 dialogues: 14 involving an American English customer and 14 involving a German customer; all dialogues involved Italian agents. Each group consisted of 7 SO and 7 MM dialogues.

For each interaction, each site (CMU; UKA; ITC-Irst) recorded an audio file containing the original voice of the local speaker and the other party's translated message. This produced 56 audio files: 28 for the agent's side (agent's original speech and customer's translated voice), 14 for German and 14 for English on the customer sides (customer's original speech and agent's translated voice). The dialogue corpus consists of 16.5 hours of dialogue length: 8.5 hours of English-Italian, 8 hours for German-Italian. The average duration of dialogues is 35 minutes (range: 19-59 minutes).

The audio files were transcribed in accordance to the VERBMOBIL conventions, using the TransEdit annotation tool. Besides orthographic words, the transcription files contain:

- annotations for spontaneous phenomena: false starts/repetitions, empty pauses, filled pauses, human noises, word interruptions and breaks, turn breaks and incomprehensible utterances. Technical interruptions are also marked;
- annotations for gestures, as three-line-comments added at the end of the corresponding turn. Reported information includes: gesture identification (progressive number, user, temporal relation with the spoken turn), gesture description (on the base of the used White Board commands) and gesture goal (selection, pointing, connection, words). Gestures were annotated using videos recorded at the Italian side. For details concerning gesture annotation conventions see (Burger, Costantini and Pianesi, 2002).

The two halves of each dialogue transcription (containing annotations) were aligned, in order to compare original and translated turns with their replies, and classify turns into *successful*, *partially successful* and *non-successful*:

Successful turns were those having good translations, from the grammatical, syntactical and semantic point of view.

Partially successful turns had poor or bad translation, either because of grammatical or syntactical errors, or

because some words were badly translated or not translated at all. At the same time, the translation managed to preserve (part of) the original message, so that the targeted party could react properly. A typical example is when the translated turn contains less information than the original turn — e.g., it contains the hotel name and the double room price, but the hotel category has been dropped. Another example of a partially translated turn is when many parts of the original utterance are omitted, but what remains still permits the other party to understand the message. E.g., the original turn sounds: "you can find a skating rink at Cavalese", and the translation is "skating Cavalese".

A turn was labeled as **non-successful** if the other party couldn't understand any component of the original utterance, or else the original utterance produced no translation, because of system errors (when the system fails to produce a translation, it issues a "no-tag" message, or a series of question marks).

"Turn repetitions' (the speaker repeats her utterance because of the system's errors) were counted as well.

6. Speech Input: Nespole!'s Multilingual and Multimodal Corpus

6.1. Turns, tokens, types

The total number of spoken turns, word-tokens and word-types (used vocabulary) were counted for each dialogue.

A turn is operationally defined as a speaker contribution between a switching-on and a switching-off of the microphone button in the NetMeeting® window of the Nespole! monitor.

A word-token is an occurrence of a given word-type — e.g., the sentences "Paul is the brother of John" and "John is the brother of Paul" contain 12 word-tokens and 6 word-types.

We obtained an average number of 73 turns per dialogue, 37 from agents and 36 from customers (39 for German customers and 33 for English customers); given that each dialogue lasted 35 minutes on average, the time lag between two consecutive turns is 30 seconds (average dialogue length in seconds divided by number of turns). That time span includes: the time during which the first turn is spoken, the translation time (including delays due to the network) and the time during which the translated message is uttered at the other site. Since turns are very brief (6.98 tokens on average for agents and 6,56 for customers) most of the time was 'waiting' time.

The average number of word-tokens uttered by the speakers during each dialogue is 258 for Italian agents (28 dialogues), 254 for German customers (14 dialogues) and 218 for English customers (14 dialogues). The number of word-types is 101 for agents, 103 for German customers and 82 for English customers.

By dividing the number of tokens by the number of types, we obtain the average token/type rate, which is 2.56 for agents, 2.47 for German customers and 2.66 for English customers; those values indicate how many words were uttered before a new word was introduced.

Average values and variance of all measures are similar across agents and customers and across the two conditions (Language and modality). ANOVA tests (p=0.05) ran on the number of turns, agents and customers separately, did not produce significant results. Thence, there is no evidence that modality or language affected the number of words spoken.

	Italian agent	German cust.	English cust.
turns per dialogue	37	39	33
tokens per dialogue	258	254	218
types per dialogue	101	103	82
tokens per turn	6.98	6.50	6.60
token/type ratio	2.56	2.47	2.66

Table 1: Average number of turns, tokens, types, plus rates, for each language.

6.2. Disfluences

As mentioned above, some classes of spontaneous phenomena were annotated on transcription files: a-grammatical phrases (repetitions, corrections, false starts), empty pauses, filled pauses, human noises, word interruptions and breaks, incomprehensible utterances, technical interruptions, and turn breaks; see (Burger, Costantini and Pianesi 2002) for details. For each class of spontaneous phenomena the percentage with respect to the total number of word tokens was calculated. The average percentages are very low: for seven of the eight classes they are always smaller than 3% (in most of these classes even smaller than 1%). Only the percentage of empty pauses at the customer site is a bit higher, ranging from 6% to 10%.

Spontaneous phenomena were further clustered into two groups. The first includes: empty pauses, filled pauses, human noise, and incomprehensible utterances; the second includes: word interruptions/breaks, turn breaks, a-grammatical phrases. This grouping was motivated by the hypothesis that the various disfluences have different effects on turn fluency. Specifically, pauses are expected to be less disturbing than a-grammatical phrases and turn or word breaks. This led to assigning different weights to the two groups: weight 1 to pauses and incomprehensible phrases and weight 2 to the second group. We then computed a *turn-fluency* score, as the weighted sum of the average frequencies for each class. Notice that the score did not include technical breaks because they are related to system features and hence do not inform about speech disfluences. In addition, empty pauses were not included because they were not uniformly annotated across languages. In particular, Italian annotations do not report pauses exceeding a given threshold (600 ms).

We obtained an average fluency-score of 1.27 for customers (all groups, SD = 1.15) and 1.06 for agents (all groups, SD = 1.48). ANOVA tests (p=0.05) run on customers and agents separately didn't detect any effect of modality and/or language on the turn-fluency score. Hence, there is no evidence suggesting that turn fluency is affected by the experimental condition (MM and SO) or by customer's Language (English or German).

6.3. Turn successfulness

The aligned transcription files (see §5) made it possible to compare original and translated turns with their replies, and classify original turns into successful, partially successful and non-successful. The table below reports average percentages for turns. Percentages of each class of

turns are very similar across Languages and experimental condition.

	Eng. SO	Eng. MM	Ger. SO	Ger. MM
Successful turns	28%	29%	31%	28%
Partially successful turns	32%	39%	28%	31%
Non-successful turns	39%	32%	41%	41%

Table 2: Percentages of successful, partially successful and non-successful turns for each modality and Language

6.4. Turn repetitions

The percentage of repeated turns over genuine turns is 17. The figures for each group are very similar for speech-only and multimodal conditions. In addition, no relevant differences were found between English and German dialogues.

turns	Eng SO	Eng MM	Ger SO	Ger MM
repeated turns	16%	16%	20%	18%
repetitions	34%	28%	36%	36%
other turns	50%	56%	44%	46%

Table 3: Percentages of repeated turns, repetitions and other turns for each modality and language

Each repeated turn was repeated, on average, 2 times. The diagram below shows the percentages of repeated turns and of repetitions of spoken turns for all groups; the counting of repetitions does not include the first representation of the turn and refers to both *immediate repetition* — i.e., directly following repetitions — and *delayed repetitions* — i.e., later repetitions. The third category (other) includes turns that were neither *repeated turns* nor *repetitions* of previously uttered turns; these are those turns produced only once.

6.5. Dialogue fluency

During the dialogue the speakers sometimes returned to previously discussed topics. When frequent, those *returns* complicate the dialogue flow and decrease dialogue fluency.

Returns are usually related to difficulties in successfully closing a dialogue segment. For instance, the customer does not manage to obtain clear answers to her questions, so she (temporarily) abandons the current topic and returns to it later on, asking for further clarifications. We hypothesized that multimodality positively affect dialogue fluency, as it might help speakers successfully close a dialogue segment, thus lowering their need to reiterate old topics. Hence, we expected a lower number of returns in MM than in SO. Moreover, it is also expected that this advantage should be clearer for dialogue segments dealing with spatial information, because MM provides alternative methods of conveying information about cartographic landmarks (e.g. drawings, pointing, etc).

The average number of returns per dialogue is 3.6. A rate for returns was also computed, by dividing the number of turns by the number of returns. This rate (*returns rate*) indicates how many turns were spoken in average from one return to the following, and can be used as an index of dialogue fluency: the greater the index, the better the fluency. We computed two such indices: the first over all the turns of a dialogue, and the second limited to the turns conveying spatial information. Average figures for each combination of language and modality are reported below:

	all turns		spatial turns	
MODE	SO	MM	SO	MM
German	21	24	13	11
English	19	31	15	44

Table 4: Returns rate for all turns, and for turns conveying spatial information

German dialogues have similar return rates in SO and MM conditions, both in the all-turns condition and in the spatial-information-only modality. In English dialogues the return rate is clearly higher in MM condition than in SO condition. In all-turns, we have19 turns spoken on average from one returns to the following in SO, and 31 in MM. In the only-spatial-turn condition, the figures are 15 in SO and 44 in MM. It can be concluded, therefore, that tendency for MM to be superior to SO in terms of dialogue fluency is confirmed by English dialogue, and that is especially true when spatial information is conveyed.

7. Pen-based input

7.1. Gestures frequencies

Table 9 reports the average values per dialogue for each class of MM drawings. We counted the number of *selection*, *pointing* and *connection gestures*, no matter whether they were performed using the *free-hand*, the *line* or the *elliptical/rectangular* selection function of the White Board. In addition we counted the number of times the agents used the *free hand* modality to write some words on the map; most of these being hotel or town names used in association with selection or pointing gestures.

selection	4.7
pointing	1.4
connection	1.0
words	0.5
SUM	7.6

Table 5: Average number of drawing gestures per dialogue (MM condition)

The average number of *drawing gestures* per dialogue (MM condition) is 7.6. Given that the average number of turns per dialogue is 73, this means that gestures were performed on average every 10 turns. Considering that some gestures were performed together to convey a unique meaning, the number of "meaningful" gestures (sequences) is even lower, e.g. most of the *pointing gestures* were combined with *selection gestures* emphasizing the latter, rather than conveying additional information — e.g., an area was first selected and immediately after it was "pointed" at. Counting the number of *pointing gestures* that are performed in isolation — i.e., not in association with *selection gestures* — we obtain an average number of performed gestures

per dialogue of 6.4 instead of 7.6. Such low ratios are not unexpected, in view of the fact that interaction involving spatial information was confined to a few dialogue segments.

Figures in the table are not separated for agents and clients, because the agents performed almost all the gestures (only two drawings were performed by customers, 1.9 % of the total number).

Finally, the table shows a clear preference among drawings for area selections, which resulted in 62% of the total number of drawings.

7.2. Speech-Gestures Association

Three classes of temporal integration patterns between gestures and speech were annotated: *immediately before*, *during* or *immediately after* the corresponding speech turn. The result is that almost all gestures followed the speech.

The typical sequence occurring when an agent wanted to load maps or web pages, or to perform drawings, consisted of some kind of verbal anticipation of her intentions — e.g.: "I'm going to send you a map", or "I'll show you the ice skating rink on the map" — followed by switching off the microphone, and then by gesture performance. It can be argued that this particular sequence was motivated by the push-to-talk procedure, which disfavored sequences where map-loading is performed before switching off the microphone, because the cognitive/emotional load associated with the latter assigned a higher priority to it. However, we have no independent evidence to this effect.

Other factors that might have had a role in determining the favored sequence are the time needed to select White Board commands, and the very long time needed to have the maps transferred at the other side (from one to two minutes, depending on network traffic). Given those time lags, it is conceivable that agents might have wanted to first make clear their intentions, alerting the customer that she would have to wait for a while before those intentions could be accomplished.

Few or no deictics were used. "Here" was sometimes used by the customer to inform the agent that the map, or the web page, was on his/her screen — e.g., "the map is here". No other relevant usage of deictics has been found, and even for cases like the mentioned ones, agents preferred to use locating phrases that relied only on visually available cues — e.g., "the skating rink is at the bottom right of the map", "I'm selecting it with the red color".

Those findings too seem related to the push-to-talk procedure and to the time necessary to transfer gestures. As already mentioned, users tended to avoid mixing gestures and speech: they uttered something, then switched off the microphone and finally performed the gesture. As a consequence, there was always a certain time lag between speech and gestures. Deictics, on the other hand, consist of (almost) simultaneous linguistic markers and demonstrations (gestures, in our case). In the described situation, deictics turn out to be infelicitous. Hence they were scarcely used.

8. Other results

Both the SO and the MM versions of the system were effective for goal completion: 86% of the users were able to complete the task's goal by choosing a hotel meeting the pre-specified budget and location constraints. This demonstrates that the system is sufficiently adequate for novice users to accomplish the given task with minimal written instructions, a very short initial training on using the White Board, and no further assistance during the interaction.

Qualitative data analysis shows that MM exhibited fewer ambiguous utterances. Furthermore, the ambiguities in MM conditions were often immediately solved by resorting to MM resources. This was not the case in SO, where ambiguities or mis-understood utterances often remained unresolved. In this connection it is worth spending few words commenting a typical case of unsolved ambiguities in the SO condition. It occurred with a certain frequency that while talking about a particular location — e.g., Cavalese — the necessity arose to switch the attention to a different place — e.g., Panchià. This often caused the customer to mistake Panchià for something other than a town — e.g., an hotel — thinking that the conversation still continued to focus on Cavalese. Upon realizing this, the agent usually tried to help by uttering something like 'This town is not Cavalese, it is Panchià', whose common translation was 'Panchià, not Cavalese'. The translation was not effective, however, to help the customer recover from her misunderstanding, so the ambiguity persisted, often without being solved by the end of the session. We hypothesise that the customer was unable to recover from the agent's utterance the intended meaning because the uttered translation lacked the appropriate contrastive accent, which would have been necessary to convey that Panchià is another place, and that attention should be shifted from one place (Cavalese) to another (Panchià). This, in turn, was due to the fact that NESPOLE! does not address contrastive information at the prosodic level.

As noticed, the same misunderstanding could be immediately solved in the MM condition: all the agent had to do was to select the two locations on the map, this way showing that they were different. If our hypothesis is correct, these differences are evidence in favour of the conclusion that multimodal communication can actually provide effective means to overcome HLT limitations.

As to usability, a questionnaire (System Usability Scale, developed at Digital Equipment Co. Ltd, Reading, UK) did not reveal any significant difference among users across the SO and MM conditions. When explicitly asked to choose between the MM and SO system (the question was "if you were to participate in a new experimental dialogue, which condition would you prefer?"), however, the users who acted as agents indicated a clear preference for the MM system.

9. Conclusions

Considering all the above-mentioned results, we can therefore conclude that:

- multimodal interaction is better that speech-only one when spatial information is conveyed; in particular it helps decreasing ambiguities and solving misunderstandings, and provides for a better dialogue fluency;
- multimodal input is preferred by users who were confronted with both MM and SO interaction modalities;

- influence of multimodal input on the linguistic features was not detected (no evidence of a clear effect of multimodality on linguistic variables and on the fluency score);
- influence of multimodal input on the global interaction features, including language global features, was not detected.

The fact that the study was carried on by using a real system prototype instead of, for example, through the WoZ, is of primary importance to assess the results. The 'real' system caused many errors and failures during the interactions, which, in turn, resulted in a high variability of measured variables, thus lowering the power of the statistical tests. It is remarkable that, despite these adverse conditions, the task was completed in the great majority of cases, hence proving the effectiveness of the system in supporting users, both in SO and in MM.

It should be mentioned that we didn't do more on the multimodal side — e.g., by allowing more room to multimodal interaction — because the translation modules available at the time the experiment was performed would have not been capable of providing enough support to this end. The task we devised was the best compromise between the system's capabilities at that time, and the need to provide for true pen-based gestures. Finally, most of the negative, or absent evidence we pointed out concern measures at the global level (whole dialogues). We are currently restricting our consideration to spatial segments. Hopefully, such a closer and more detailed analysis will highlight effects that are lost or blurred at the global level.

Concerning the Nespole!'s Multilingual and Multimodal corpus, it is (to our knowledge) a unique source of information for those interested in the topic of multilingual and multimodal interaction in a realistic scenario. It provides detailed annotation of the dialogues, and video-recordings of each experimental session. There are three different languages involved. Importantly, the prototype system used in the experiment challenged the users in overcoming errors made by machine. They had to find solutions with limited means. So, the corpus is a valuable source of information for studying the actual strategies users deploy in real machine-mediated communication.

The NESPOLE! consortium has already used observations and insights gathered from the corpus to improve the system — e.g., by providing users with better support for feedback about the success of the different stages of the translation process. The dialogues and behavior of the subjects involved in the experiment deserve further investigation and will serve for additional improvements of the NESPOLE! translation system and systems of similar kind.

10. Acknowledgments

The work described in this paper has been partially supported by National Science Foundation under Grant number 9982227, and by the European Union under Contract number 1999-11562 as part of the joint EU/NSF MLIAM research initiative. Any opinion, suggestion and recommendation expressed in this paper are those of the authors and do not necessarily reflect the views of the EU or of the NSF.

The authors wish to acknowledge the contribution and support by the other participants in NESPOLE! to the preparation and realisation of the experiment described here. In particular, Alon Lavie (CMU), John McDonough (University of Karlsruhe), and Loredana Taddei (Aethra).

11. References

Burger, S., L. Besacier, P. Coletti, F. Metze and C. Morel, 2001. The NESPOLE! VoIP Dialogue Database. In *Proceedings of Eurospeech 2001*. Aalborg, Denmark.

Burger, Costantini and Pianesi, 2002. *NESPOLE! Deliverable D5 - Study on Multimodality, part 1*. In NESPOLE! Project website: http://nespole.itc.it.

Lavie, A., C. Langley, A. Waibel, F. Pianesi, G. Lazzari, P. Coletti, L. Taddei and F. Calducci, 2001. Architecture and Design Considerations in NESPOLE!: a Speech Translation System for E-commerce Application. In *Proceedings of HLT2001*. San Diego, Ca.

Lavie, A., F. Metze, F. Pianesi et al., 2002. Enhancing the Usabilità and Performance of NESPOLE! – a Real-World Speech-to-Speech Translation System. In *Proceedings of HLT2002*. San Diego, Ca.

Lazzari G., 2000. Spoken translation: challenges and opportunities. In *Proceedings of ICSLP'2000*, Beijing, China.

Metze, F., J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schutz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana and E. Pianta, 2002. The NESPOLE! Speech-to-Speech Translation System. In *Proceedings HLT2002*. San Diego, Ca.

Oviatt, S.L., 1997a. Multimodal interactive maps: Designing for human performance. In *Human-Computer Interaction*, 93-129 (special issue on "Multimodal interfaces").

Oviatt, S. L., DeAngeli, A. & Kuhn, K., 1997b. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*. New York, ACM Press, 415-422.

Oviatt, S. L., 1999. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of Conference on Human Factors in Computing Systems CHI '99*. New York, N.Y.: ACM Press, 1999, 576-583.