COMPARATIVE EVALUATION OF SPOKEN CORPORA ACQUIRED BY PRESENTATION OF VISUAL SCENARIOS AND TEXTUAL DESCRIPTIONS

D. Aiello[1], C. Delogu[1], R. De Mori[2], A. Di Carlo[1], M. Nisi[1], S. Tummeacciu[1]
[1] Fondazione Ugo Bordoni, v. B. Castiglione 59, 00142 Roma, Italy
e-mail: demetrio;cristina;adicarlo@fub.it
[2] Université d'Avignon et des Pays de Vaucluse, BP 1228, 84911 Avignon Cedex 9
e-mail: renato.demori@lia.univ-avignon.fr

ABSTRACT

The paper describes a system, in JAVA, for written and visual scenario generation used to collect speech corpora in the framework of a Tourism Information System.

Experimental evidence shows that the corpus generated with visual scenarios has a higher perplexity and a richer vocabulary than the corpus generated using the same conceptual derivations to produce textual scenarios. Furthermore, there is evidence that textual scenarios influence speakers in the choice of the lexicon used to express the concepts more than visual scenarios.

1. INTRODUCTION

This paper deals with the evaluation of different methods for collecting corpora of spoken signals. These corpora are essential for estimating the parameters of various models used for automatic speech recognition and understanding (ASRU).

For applications other than dictation, e.g. for spoken database query or for speech-to-speech translation, corpora should contain spontaneous speech elicited from speakers without presenting them a text they should just read. Certainly, some information has to be presented to the speaker about a given topic. Nevertheless, presentations should allow for a certain freedom in the choice of words and sentences

In practice, *situation* descriptions are created to act as constraints on what speakers will say by choosing their proper words and sentences. Situations are descriptions of an aspect of the application domain.

The question addressed in this paper is: in what form and what information should be presented to a speaker in order to achieve a rich combination of pertinence and variety in the acquired corpus? Types of presentation described in the following are textual and visual.

Textual presentations consist in texts describing user situations. Analogously visual presentations consist in displayed scenes showing user situations. Speakers, acting as users, have to suppose to be in one of these situations; they have to say something appropriate to the described situation and consistent with the objectives of the application which has been described before the acquisition sections.

As opposed to the case of newspaper dictation, texts or scenes that have to inspire speakers have to be generated by a procedure that respects the constraints imposed by the application. If such a generation is made by computers, it is convenient to follow a generation paradigm based on a formal method. Such a method should generate a sort of conceptual representation from which *sentence contents* are derived in visual or textual description form.

In order to elicit spontaneous sentences for spoken corpora collection, the use of scenarios is crucial. A *scenario* can be defined as a description of an actual task that each speaker has to accomplish. Scenarios should stimulate the subjects to generate sentences with a large variety of words and language constructs. Until now, textual scenarios (TS) have been frequently used for corpus acquisition. The limit of these scenarios is that they are likely to influence a speaker in the choice of the words used to express the concepts. So far, there isn't a well-known method for designing scenarios avoiding the linguistic bias introduced by textual scenarios. Table-scenarios [3], or inserted graphic representation in textual scenarios [4] have also been used.

As the objective is that of producing a rich training corpus with the greatest variability of sentences expressing domain conceptual structures, it is important to evaluate corpora acquired with textual or visual presentations or a combination of both. A written text presentation is likely to influence a speaker in the choice of the words used to express the concepts, while visual scenarios (VS) may inspire a greater variety of sentences, even if understanding computer graphics may be more difficult than understanding text.

This paper describes a working system, written in JAVA, for generating sentence contents in textual and visual forms for Traveler Domain tasks of the EUTRANS project for speech-to-speech translation [5]. Statistical language models built with the two approaches have been evaluated in terms of perplexity and vocabulary size. Details of the evaluation are given in section 4. The generation of conceptual representations is described is section 2. The generation of situation representations is described is section 3.

2. GENERATION OF CONCEPTUAL REPRESENTATIONS

It is useful, in practice, to consider scenario generation as a sequence of two processes. The task of the first process is to produce a formal description of the expected spoken message contents, while the second process has to compose a text or a display image corresponding to the description. In the case of images, the second process can be further subdivided into two further steps. The first one is a generator of the visual components that should appear in the scene, together with their logical relations. The second one has to assemble the scene elements in such a way that they satisfy the constraints imposed by logical relations and those imposed by geometric consistency. An example of a logical constraint is that a bed in a hotel room should lie on the floor, while an example of geometric constraint is that two physical objects cannot share the same physical space.

Details of the generation using a generative grammar of frame data structures are given in [1].

A frame has a frame header and (attribute, value) pairs. The values of pairs can be nonterminal symbols to be further expanded into frame structures. In some cases, the grammar generates only the attribute part of the pair and lets the human subject use her/his fantasy to decide the values.

The start symbol of the grammar σ represents the class of all possible scenarios.

The first rule is of the type:

$$\sigma \rightarrow FR1 \mid FR2 \mid FR3 \mid .....$$

where FRi are non-terminal symbols from which frame structures are generated. For example, FR1 is rewritten into the following frame structure:

The structure for a HOTEL_ACTION is a collection of attributes, *actor, action, place* and corresponding values which are, in this case, nonterminals generating other frame structures.

$$FR1 \rightarrow \begin{vmatrix} HOTEL\_ACTION \\ actor \quad FRACTOR \\ action \; FRACTION \\ place \quad FRPLACE \end{vmatrix}$$

FRACTION is a simple non-terminal that can be rewritten into non-terminals generating frames representing a request for information, a cancellation of a reservation, a complaint, a modification.

FRACTOR generates frames about the purpose of the intervention and the other human subjects (himself, wife, friends, etc.) for which the action is performed.

FRPLACE generates frames describing hotels, airports, railway stations etc.

A *derivation* of the grammar is a complete hierarchy of frames which no longer contain non-terminal symbols.

Functions are associated to grammar rules. They provide constraint satisfaction and assemble in a coherent way the components of a situation description.

Rewriting rules and associated functions follow a taxonomy, the main components of it are described in the following using a simplified version of the frame grammar.

FRACTOR in the considered application describes a person who calls the Tourist Information System on behalf of himself, his family or other people. This is represented in the taxonomy by the following expression:

$$FRACTOR \rightarrow FRC \mid .....$$

$$FRC \rightarrow \begin{vmatrix} CALLER \\ on\_behalf\_of \; \{HIMSELF \mid FAMILY \mid OTHERS\} \end{vmatrix}$$

The symbol | indicates logical disjunction.

FRACTION→FRINFO | FRREQUEST | FRCANCELLATION | FRCOMPLAINT | FRMODIF}

$$FRINFO \rightarrow \begin{vmatrix} INFORMATION \\ information\_type\{INFO\_ROOM \mid INFO\_HOTEL \\ \qquad\qquad \mid INFO\_INFO\_SERVICE\} \\ location\{ROOM \mid OTHER\_PLACE\} \\ time\{.....\} \end{vmatrix}$$

$$FRCANCELLATION \rightarrow \begin{vmatrix} CANCELLATION \\ location\{ROOM \mid OTHER\_PLACE\} \\ time\{.....\} \\ object\{ROOM \mid SERVICE\} \end{vmatrix}$$

This is obviously just a fragment of the complete grammar describing the application taxonomy. It shows how the same fragment can be used as a filler (value) of slots (arguments) of different frames. Recursion takes place when a frame is filler of a slot of itself.

A method of the JAVA class *Syntax* randomly selects and apply rewriting rules and the associated functions to generate a symbolic representation of a situation.

A natural language description of the scenario is then generated by rules starting from the conceptual structure represented by the result of the derivation.

3. GENERATION OF SITUATION REPRESENTATIONS

Every terminal symbol of the grammar is associated with a text *generation function*. In each of them, there are two kinds of statements: first-type statement can directly generate and display one or more phrases in a Natural Language; second-type statements are able to set the display position for phrases generated and displayed by some other *generation functions* recursively called.

In order to produce a graphical image, another set of rules is used to generate *instances of JAVA* classes from the specific grammar derivation. Specific constraints can also be obtained by these rules.

Scenarios are generated by the methods of a class *DrawScenario* with the help of a *layout manager*, a classical interface component (JAVA has a number of them) that receives image components and their relations and produces the 2-D image by setting size and position of each image.

This component is still driven by a constraint satisfaction algorithm that is based on geometric constraints.

4. EVALUATION

With the aim of exploring the possible difference between textual and visual scenarios, an evaluation test has been performed with 100 subjects subdivided into two groups. Textual scenarios were presented to speakers in group-TS while speakers in group-VS were exposed to visual scenarios.

In order to verify the following hypotheses: (i) sentences obtained by VS are more complex than those obtained by TS; and (ii) sentences obtained by VS are more difficult to model than sentences obtained by VS; perplexities were computed using the CMU-Cambridge Statistical Language Modeling Toolkit v2 [2].

A train set of 400 sentences was used (200 visual and 200 textual) with 8515 words and a vocabulary of 898 type. The test set for the VS group was made of 78 sentences (with 1594 words), while the test set for the TS group was made of 78 sentences (with 2119 words).

Table 1 shows the perplexity computation for both the trigram and the bigram models on the sentences produced starting from visual scenarios (Visual Corpus) and those produced from textual scenarios (Textual Corpus).

TRIGRAMS MODEL (a)

	Perplexity	Out-of-Vocabulary
Visual Corpus	36.92	120
Textual Corpus	20.73	77

BIGRAMS MODEL (b)

	Perplexity	Out-of-Vocabulary
Visual Corpus	41.44	120
Textual Corpus	26.05	77

Table 1: Perplexity values and out-of-vocabulary words obtained with the trigram model (a) and the bigram model (b).

As for each derived conceptual representation a textual scenario and a visual scenario were generated, the corpora obtained with the two generation mechanisms are directly comparable even if the size of the corpus is not big. The differences in perplexities of the test sets are consistently and substantially higher for the language generated with visual scenarios for different types (bigram and trigram) of language models.

Furthermore, the number of Out-Of-Vocabulary (OOV) words is much higher (120 vs. 77 for a vocabulary of 898 words) for the corpus obtained with visual scenarios with respect to the corpus obtained with text scenarios.

An interesting question concerns the similarity between the two corpora. In other words, does the textual corpus look like a subset of the visual corpus or it is a substantially different corpus? In the latter case, sentences acquired with the two approaches can be merged into a compound corpus much richer than each component.

In order to answer this question, two LM were built. One, called LMT was built using all the sentences of the textual corpus and the other, called LMV was built using all the sentences of the visual corpus.

The following perplexities were then computed PPTV is the perplexity of the visual corpus obtained using the LMT of the textual corpus and PPVT is the perplexity of the textual corpus obtained with the LMV of the visual corpus. Bigram and trigram LM were considered. The results are shown in Table 2.

	PPTV	PPVT
bigram LM	66.87	55.37
trigram LM	64.07	54.96

Table 2: perplexities of visual (PPTV) and textual scenarios (PPVT) using LMT and LMV respectively.

Let us define OOVTV as the percentage of OOV in visual scenarios with respect to the vocabulary of LMT and OOVVT as the percentage of OOV in textual scenarios with respect to the vocabulary of LMV. These percentages are shown in table 3.

OOVTV %	OOVVT %
16.79	11.76

Table 3: Percentage of OOV in visual scenarios with respect to the vocabulary of LMT (OOVTV) and percentage of OOV in textual scenarios with respect to the vocabulary of LMV (OOVVT).

The results show that the two corpora are substantially different even if each sentence in a corpus has a corresponding sentence in the other corpus which has been elicited with the same conceptual structure.

The union of the two corpora can then be used to train a LM for generating word hypotheses. The sentences of the compound corpus have been translated in the languages of the application and use to train automata for automatic translation.

An analysis of word intersection was performed by extracting from each sentence of TS the *key-phrases* corresponding to frame names. For example, in the sentence:

You are calling the reception of the Hilton hotel, to know how much is the cost of room service

the key-phrase is <the cost of room service> which corresponds of the attribute ROOM_SERVICE of the frame COST which is a possible value of the attribute ¨content¨ of INFO_INFO_SERVICE.

All the key-phrases (often made of just one word) were searched within the two corpora TS and VS. It was observed that the sentences in TS contained almost all the words of key_phrases (key_words), which on the contrary were rarely found in the sentences of VS. Table 4 summarizes the result for a sample of 261 sentences generated after the presentation of 29 scenarios.

Number of sentences	Key-words in TS	Key-words in VS
261	1253	483

Table 4 : Results of key_word intersection analysis

This result suggests that when subjects were exposed to VS, they used words that are not frame labels. On the other hand, a synonym analysis in the two corpora showed that the VS corpus contains an average of 7 synonyms per word, while only 3 are in the TS corpus. On average, 2 synonyms are shared between the two corpora.

In order to evaluate the comprehension of visual scenarios, a criterion has been defined to evaluate the comprehension correctness of all the sentences in VS. In particular, for each scenario, two key-words are defined: one is *essential* to the comprehension and the other one is *optional*. In order to reach a *narrow* comprehension, both key-words have to be found in the sentence. If only the essential key-word is found, it is assumed that comprehension has been *broad*. If the essential key-word is not found, then it is assumed that the subject did not understand the scenario.

Evaluation of the degree of understanding has been performed on 381 sentences from 28 visual scenarios: 90% of sentences obtained *narrow* comprehension; 7% *broad* comprehension; and 3% obtained no comprehension.

The overall results of scenarios evaluation confirmed the conjecture that textual scenarios influence speakers in the choice of the lexicon used to express the concepts more than visual scenarios. Furthermore, they also showed that the sentences in VS have very high lexical differentiation.

These results can be explained by considering the role of language in cognition and the social psychology of experiments with human subjects. Language helps to segment and categorize reality. Dealing with an already linguistically described situation saves the effort of finding the appropriate segmentation and categorization of the situation, but, at the same time, it does not encourage to find alternative segmentations and categorizations. Furthermore, the experiment is a specific social context which tends to induce an attitude of deference with respect to the experimenter and the experimental material in human subjects. Both factors result in a tendency of the subjects who are exposed to linguistically described situations to react to these situations by using the same restricted language which is used in the experimental materials offered to them. On the other hand, subjects who are exposed to visual situations must find their own way to find linguistic descriptions of visual material and are not influenced to rely on the experimenter's language.

5. CONCLUSIONS

The system described in this paper has demonstrated the utility of using conceptual grammars for generating textual and visual scenarios. Furthermore, graphic objects (e.g. instances of JAVA classes) can be easily generated from a frame description language.

As a domain taxonomy is explicitly represented in the language, statistics of the derivations represent the coverage of the generated corpus.

Experimental evidence shows that the corpus generated with visual scenarios has a higher perplexity and a richer vocabulary than a corpus generated using the same conceptual derivations to produce textual scenarios. Furthermore, there is evidence that textual scenarios influence speakers in the choice of the lexicon used to express the concepts more than visual scenarios.

Corpora generated with textual and visual scenarios using the same derived conceptual descriptions are substantially different and can be advantageously combined into a richer corpus to be used to train an LM and translation automata.

6. REFERENCES

[1] Aiello D., Delogu C., De Mori R., Di Carlo A., Nisi M. and Tummeacciu S. "Automatic Generation Of Visual Scenarios For Spoken Corpora Acquisition." *Proceedings of Int. Conf. of Spoken Language Processing, Sydney, 1998*

[2] Clarkson P., Rosenfeld R. "Statistical Language Modeling Using the CMU-Cambridge Toolkit." *Proceedings of Eurospeech'97, Rhodes 1997, vol. 5, pp. 2707-2710*

[3] Delogu C., Di Carlo A., Sementina C., Stecconi S. "A Methodology for Evaluating HumanMachine Spoken Language Interaction.", *Proceedings of Eurospeech'93, Berlin 1993,Vol. 2, pp. 1427--1430.*

[4] Dybkjaer L., Bernsen N.L., Dybkjaer H. "Scenario Design for Spoken Language Systems Development." *Proceedings of ESCA Workshop on Spoken Dialogue Systems, Vigso 1995, pp. 93—96*

[5] http://hermes.zeres.de/Eutrans/eutrans.html