

Results of the French Evalda-Media evaluation campaign for literal understanding

H Bonneau-Maynard, C Ayache, F Bechet, Alexandre Denis, A Kuhn, F Lefevre, D Mostefa, M Quignard, S Rosset, Christophe Servan, et al.

▶ **To cite this version:**

H Bonneau-Maynard, C Ayache, F Bechet, Alexandre Denis, A Kuhn, et al.. Results of the French Evalda-Media evaluation campaign for literal understanding. The fifth international conference on Language Resources and Evaluation (LREC 2006), May 2006, Genes, Italy. hal-01160167

HAL Id: hal-01160167
https://hal.science/hal-01160167

Submitted on 4 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Results of the French Evalda-Media evaluation campaign for literal understanding

H. Bonneau-Maynard[1], C. Ayache[2], F. Bechet[3], A. Denis[4], A. Kuhn[2], F. Lefevre[1,3], D. Mostefa[2], M. Quignard[4], S. Rosset[1], C. Servan[3], J. Villaneau[5]

(1) LIMSI/CNRS, (2) ELDA, (3) LIA/UAPV, (4) LORIA, (5) VALORIA

Abstract

The aim of the MEDIA-EVALDA project is to evaluate the understanding capabilities of dialog systems. This paper presents the MEDIA protocol for speech understanding evaluation and describes the results of the June 2005 literal evaluation campaign. Five systems, both symbolic or corpus-based participated to the evaluation which is based on a common semantic representation. Different scorings have been performed on the system results. The understanding error rate, for the *Full* scoring is, depending on the systems, from 29% to 41.3%. A diagnosis analysis of these results is proposed.

1. Introduction

Various influential projects have built the foundations of evaluation methodologies for spoken dialog systems, such as the ATIS (MADCOW, 1992) and COMMUNICATOR (Walker et al., 2002) projects in the USA, and the European project DISC (Dybkjaer and Bernsen, 1998). The dynamic and interactive nature of dialog makes it difficult to build a reference corpus of dialogs against which systems can be evaluated.

The aim of the MEDIA-EVALDA evaluation campaign is to test an automatic evaluation methodology for man-machine dialog systems. The evaluation methodology is based on a paradigm that uses test sets taken from a corpus of real-world dialogs, a semantic representation of dialog and common evaluation metrics. The evaluation environment relies on the assumption that, at least for database query dialog systems, it is possible to define a common semantic representation to which each system can convert its internal representation. A protocol has been designed to test the understanding capacity of dialog systems, in both literal and contextual mode. Systems from both academic organizations and industrial sites were involved in the project. ELDA coordinates the project and LIMSI acts as scientific supervisor.

This paper presents the protocol and results of the June 2005 literal understanding campaign. The common semantic representation is first defined in section 2.. Then section 3. gives a brief description of each system taking part in the campaign. The evaluation protocol is then given in section 2. Finally, the results of the June 2005 campaign, are given and analyzed in section 4.2.

2. Semantic representation

The speech understanding module is the front end of the dialog manager. Its role is to analyze the user query and to produce a representation of its semantic content that allows the dialog manager to take a decision about the dialog follow-up taking into account the context. The task chosen is hotel room reservation, with touristic information as an additional topic of the dialog.

The MEDIA evaluation paradigm relies on a common generic semantic representation (Bonneau-Maynard and others, 2005). The representation is based on an attribute-value structure in which conceptual relationships are implicitly represented by the name of the attributes.

The semantic representation relies on a hierarchy of basic attributes, which are identified in a semantic dictionary, jointly developed by the MEDIA consortium. This conceptual hierarchy provides also a set of relationships between semantic units. A dialog consists of a number of turns. Each turn of a dialog is segmented into one or more dialogic segments and each dialogic segment is segmented into one or more semantic segments with the assumption that a semantic segment corresponds to a single attribute. An example of a semantic representation of a client utterance is given in Figure 1.

A semantic segment is represented by a 5-tuple which contains:

- the mode: affirmative '+', negative '-', interrogative '?' or optional '~ ',
- the name of the attribute representing the meaning of the sequence of words,
- the value of the attribute,
- some optional links: pointers to related segments in previous utterances (only useful for contextual semantic representation),
- an optional comment on the segment.

The order of the 5-tuples in the semantic representation follows their order in the utterance. The attribute values are either numeric units, proper names or semantic classes merging lexical units which are synonyms for the task. The modes are assigned in a per segment basis. This allows to disambiguate sentences such as *"not in Paris in Nancy"* which could otherwise be misleading for the dialog manager.

2.1. Semantic dictionary

The basic attributes can be divided in several classes. The **database attributes** correspond to the attributes of the database tables (eg `BDObject` or `payment-amount`). The **modifier attributes** (eg `comparative`) are linked to database attributes and used to modify the meaning of

word seq.	mode	attribute name	normalized value
euh	+	null	
oui	+	response	yes
l'	+	refLink-coRef	singular
hôtel	+	BDObject	hotel
dont	+	null	
le prix	+	object	payment-amount
ne dépasse pas	+	comparative-payment	less than
cent dix	+	payment-amount-integer-room	110
euros	+	payment-unit	euro

Figure 1: Example of the semantic attribute/value representation for the sentence *"hum yes the hotel which price doesn't exceed one hundred and ten euros"*. The relations between attributes are given by their order in the representation and the composed attribute names. The segments are aligned on the sentences.

the relying database attribute (eg in Figure 1 the comparative attribute, which value is less than) is associated to the payment-amount attribute). **General attributes** are also defined as command-task which includes the different actions that can be performed on objects of the task, or command-dial with values cancellation , correction ... The general attribute refLink is dedicated to the annotation of linguistic references. A connector general attribute (with values and , or , implies , explains , opposes) is also defined to represent logical links between portions of queries.
The general and modifier attributes are domain independent and were directly derived from other applications whereas most of the database attributes were derived from the database linked to the system.
The set of normalized values associated to each attribute is defined in the semantic dictionary with 3 different possible configurations: a value list (eg comparative with possible values around , less-than , maximum , minimum and more-than), regular expressions (as for dates), or open values (i.e. no restrictions, as for client names).

2.2. From flat annotation to hierarchical representation

Hierarchical semantic representation is powerful as it allows to explicitly represent relationships between segments, possibly non-adjacent in the transcription of the query. On the other hand, a flat representation facilitates the manual annotation of the data. It has then been decided for the MEDIA annotation scheme to preserve the relationships, by defining a set of **specifiers** which are combined with database or modifier attributes. For example, in Figure 1, the attribute comparative-payment is derived from the combination of the comparative attribute and the payment specifier and the attribute payment-amount-integer-room is derived from the combination of the payment-amount-integer attribute with the specifier room . The combination of the specifiers and the attribute names allows to recompose a hierarchical representation of a query from its flat annotation.

2.3. Semantic content of the MEDIA corpus

The semantic dictionary defined for the MEDIA project includes 83 basic attributes and 19 specifiers. The combination of the basic attributes and the specifiers - automatically generated by the annotation tool - results in a total of 1121 attributes that can be used during the annotation process. The 83 basic attributes include 73 database attributes, 4 modifiers, and 6 general attributes. The total number of distinct normalized values in the training set is around 2.2k.

	train	dev	test
#utterances	10965	1009	3003
average #words per utt.	4.8	5.4	6.2
number of different words	2115	794	900
number of observed attributes	29980	3125	11849
average #attributes per utt.	2.7	3.1	3.9
number of different attributes	144	106	129

Table 1: Main characteristics of the client utterances in the training, development and test corpora.

Semantic annotation has been done on the dialog transcriptions, using a the LIMSI Semantizer annotation tool[1]. Semantizer ensures that the provided annotations comply with the semantic representation defined in the semantic dictionary. An on-line verification is performed on the attribute value constraints. In order to verify their quality, periodic evaluations of the annotations were performed. The attribute inter-annotator agreement is always greater than 80%, resulting in a Kappa of more than 0.8, commonly considered as good.

Table 1 gives details on both training, development and test corpora. The most frequent attribute is the yes/no response (17%), followed by reference attributes (6.9%) and command-task (6.8%). It is interesting to note that the most frequently encountered attributes are task-independent (localization, time, ...) and that task-dependent attributes (hotel, room...) represent only 14.1% of the observed attributes. A total of 144 distinct attributes appear in the training corpus. Only one attribute of the development corpus was not observed in the training corpus.

3. System description

The five systems which have participated to the evaluation are based on different approaches. LIMSI-1 and LIA use corpus-based automatic training techniques, LORIA and VALORIA systems rely on hand-crafted symbolic approaches and LIMSI-2 system is mixed.

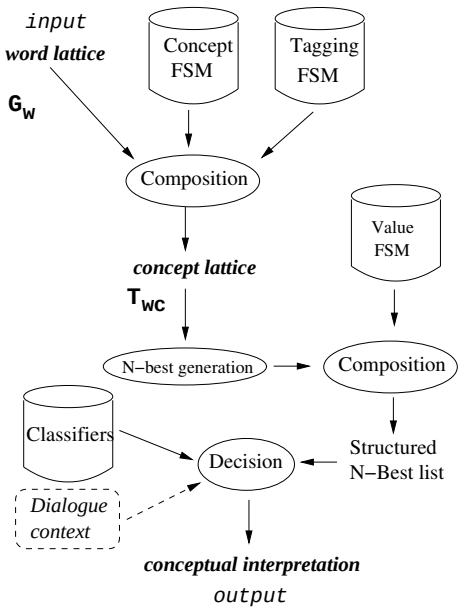


Figure 2: Overview of the LIA system

3.1. LIA system

In the Spoken Language Understanding module developed at the LIA, interpretation starts with a translation process in which stochastic Language Models are implemented by Finite State Machines (FSM) which output labels for semantic constituents. These semantic constituents are called *concept tags* and are noted γ. They correspond to the 83 concept tags defined in the MEDIA ontology (specifier and mode information is related to another interpretation level in our system). To each concept tag γ is attached the word string γ^w supporting the concept and from which the concept value (e.g. date, proper name or numerical information) can be extracted. The interpretation of an utterance containing L concepts is represented by both a *concept tag sequence* (noted $\Gamma = \{\gamma_1, \gamma_2, \dots \gamma_L\}$) and the corresponding concept word string sequence supporting each tag (noted $\Gamma^w = \{\gamma_1^w, \gamma_2^w, \dots \gamma_L^w\}$). There is an FSM for each elementary conceptual constituent. These FSMs are transducers that take words as the input and output the concept tag conveyed by the accepted phrase. They can be manually written for domain-independent conceptual constituents (e.g. dates or amounts), or data-induced for the concepts specific to the MEDIA corpus. All these transducers are grouped together into a single transducer, called *Concept FSM*, which is the union of all of them. In order to find the best sequence of concept tags for a sequence of words an HMM tagger, also encoded as an FSM is trained on the MEDIA training corpus. This FSM is called *Tagging FSM*. Finally, a last transduction process is applied to each word string γ^w in order to associate a normalized value to each concept detected; this is done with the transducer *Value FSM*. This interpretation strategy is presented in detail in (Raymond and others, 2006) and is summarized on figure 3.1.. All the operations presented on the FSMs are made with the AT&T FSM toolkit (Mohri et al., 2002). The result of the translation process is a *Structured N-Best* list of interpretations. The last step in this interpretation process consists of a decision module, based on classifiers, choosing an hypothesis in this n-best list. In this MEDIA evaluation, two classifiers have been used in order to deal with the high ambiguities of the concept tags *refLink* and *connector*.

[1] *download at http://www.limsi.fr/Individu/hbm/*

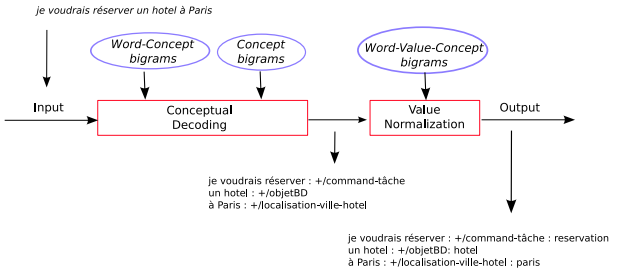


Figure 3: Overview of the LIMSI-1 system

In the results presented in table 2, the interpretation hypotheses output by our system did not include the specifier tags. This explains the huge drop in performance between the *relaxed* (no specifiers) and *Full* evaluation results.

3.2. LIMSI-1 system

The LIMSI-1 system is founded on a corpus-based stochastic formulation. The initial 2-level stochastic understanding model has been recently extended to a 2+1-level model, where an additional stochastic level is in charge of the attribute value normalization (Bonneau-Maynard and Lefevre, 2005). Figure 3 shows an overview of the LIMSI-1 system. Two stages are composed to produce the final result : a first step of conceptual decoding produces the modality and attribute sequences associated to word segments, then a final step translates the word segments into normalized values.

Basically, the understanding process consists of finding the best sequence of concepts given the sequence of words in the user query under the maximum likelihood framework. The first decoding stage aligns an attribute and its modality to each sub-sequence of the query. Bigrams of words conditioned on concepts (i.e. attribute + modality) are used during the decoding stage along with bi-grams of concepts. Some lexical classes are used to improve the generalization of the word bigrams.

In a second stage, the segmented word strings have to be converted to their expected normalized form, as given in the semantic dictionary. This normalization step was formerly obtained by means of semi-manual rules. In the LIMSI-1 system, the model has been extended with an additional level for the attribute value normalization. Due to data sparseness, a full model (i.e. with 3 embedded levels of decoding) is not straightforwardly applicable and a variant has been developed where the conceptual decoding and value normalization phases are decoupled (thus the 2+1 levels). This new model has been completed with 3 new techniques, leading to a global 20% relative improvement on the development set : penalty-based stochastic normalization, modality propagation and hierarchical recomposition.

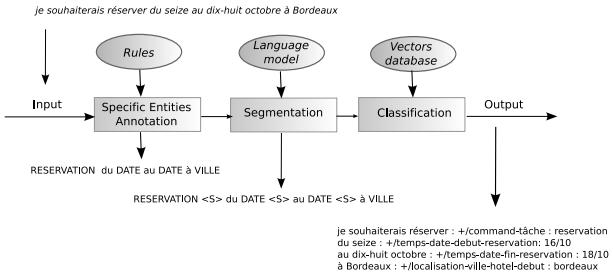


Figure 4: Overview of the LIMSI-2 system

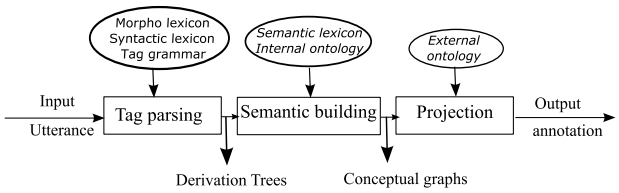


Figure 5: Overview of the LORIA system

3.3. LIMSI-2 system

The LIMSI-2 system uses the same approaches and methods used for automatic detection of dialog acts (Rosset and Tribout, 2005) and consists of three modules: specific entities detection, utterance semantic segmentation and then automatic semantic annotation. The specific entity taggers use rewriting rules which work like local grammars with specific dictionaries. They replace the specific entities by tag words expressing their types. These specific entities allow to reduce lexical variability of the utterances. Splitting an utterance into semantic units is done by inserting a boundary pseudo-word at specific places in the utterance. The insertion method uses a 4-gram language model trained on the normalized transcripts (with specific entities). We used a memory based learning approach to predict the class of these three aspects and more specifically the TIMBL implementation (Daelemans and others, 2003). MBL works by finding the vector in the training database closest to the test one. We used a Jeffrey divergence for the classification of the attribute and the value and Manhattan distance for the classification of the mode. The training corpus has been used to build the vector's models. For the classification of value, the first seven tag-words (a null word is used when needed) are used as features. For the classification of the attribute, the features are the first seven tag-words, the hypothesis value, the hypothesis attribute of the preceding semantic unit (null if needed). And for the classification of mode, the first seven tag-words, and the hypothesis value are used as features. Afterward, a post-processing is used in order to find the normalized value. Figure 4 shows an overview of the general LIMSI-2 system.

3.4. LORIA system

The approach of the LORIA system is based on deep-parsing, and description logics :

- A LTAG parser (Crabbe et al., 2003) produces a syntactic analysis. The system considers only partial derivations since it proved to be effective on spoken language. The longest derivations are kept and the others are disregarded.

- A compositional semantic builder produces a conceptual graph from the syntactic analysis. The conceptual graph is tested against an internal ontology so that inconsistent relationships are removed.

- A projection module flattens the graph and constructs the target representation format. The conceptual graph is first translated from the internal ontology into description logics formulas and each instance is tested to retrieve its most specific instantiator concepts in the external ontology. Finally each concept is written as an attribute-value pair in the MEDIA formalism and ordered along the sentence. Although the position of a concept in a given sentence does not make much sense, we kept this information from the parser, which maps trees on words and concepts on trees.

The system does not perform any training and does not need an annotated corpus but requires a high-quality annotation guideline. The system is based on hand-written resources: a morphological lexicon extracted from Multext lexicon (5,400 words, and 3,000 lemma), a syntactical lexicon created using simple heuristics (like: nouns anchor noun trees), a very small LTAG grammar (80 trees), a semantic lexicon used to produce the conceptual graphs (150 schemes), an internal ontology to check the conceptual graphs (220 concepts) and an external ontology whose concepts are defined in terms of internal concepts (130 concepts).

The advantage of the approach is that it focuses on semantic processing with a central role of the ontology and distinguishes the understanding abilities from the projection itself. But it is strongly syntax dependent and thus needs a robust grammar to parse dialog transcriptions. For more information on this system, please refer to (Denis et al., 2006).

3.5. VALORIA system

The LOGUS system implements a logical approach to the understanding of spoken French (Villaneau et al., 2004). It is relevant for a limited domain but yet much wider than the standard systems: the understanding is not frame-based but a semantic knowledge of the application domain is used.

Target language and Parsing

In the place of semantic frames, we use logical formula according to the illocutionary logic of D. Vanderveken. Concepts and conceptual structures are used in order to enable the logic formula to be convertible into a conceptual graph. The resulting graph expresses the meaning of the utterance, regardless of its linguistic form. During the parsing, constituents of the parsed sentence are gradually combined so as to join robustness and precision. As constituents increase, their meaning becomes more specific. Several different formalisms are used in sequence; they are adapted from standard syntactic formalisms in order to associate syntactic and semantic arguments. Syntactic constraints are gradually relaxed to cope with agrammaticalities.

LOGUS **and the** MEDIA **Project**

Adapting the LOGUS system to the MEDIA task was not a very difficult task: hotel reservation is a delimited and quite simple task. The main difficulties were to translate the logical formula provided by LOGUS into the MEDIA required

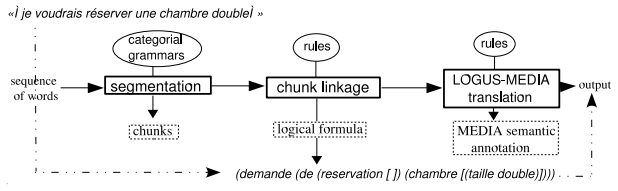


Figure 6: Overview of the LOGUS system

semantic annotation. We have chosen (perhaps mistakenly so) not to change LOGUS main content in order to have the right results for its evaluation. However, we have realized that we had to delete some rules in order to respect the order of the conceptual segments of the reference semantic annotation. With this choice, 2/3 of LOGUS errors were coming from the translation into the semantic annotation. Half of them are quite simple to correct. For the others, the logical formula LOGUS provides doesn't contain the necessary information; for example, LOGUS gives the same result for "cent euros maximum" and "maximum cent euros" as these phrases give two different MEDIA annotations.

4. Evaluation campaign

Following a dry-run in April 2005 on a 1k utterance set which enabled the definition of the test protocol, the literal evaluation campaign was performed in June 2005 on a test set of 3k utterances extracted from dialogs chosen, transcribed and randomly mixed by ELDA.

4.1. Evaluation protocol

Each participant benefited from the same semantically annotated 11K utterance training corpus to enable the adaptation of its models to the task and the domain, as well as the semantic dictionary and the annotation manual. The mean number of words per utterance in the training corpus is 4.8. The 3786 word lexicon of the MEDIA corpus of and the list of 667 values for the open-value attributes which appear in the corpus were also given to the participants. The mean number of observed attributes per utterance is 2.7. 144 different attributes were observed in the training corpus.

As observed from the inter-annotation experiment (86% agreement), manual semantic annotation of a test corpus is not a straightforward process. Some variability should be allowed in the semantic representation of a query. In a post-result adjudication phase, the participants were asked to propose either modifications or alternatives for the test set annotation. A the end a consensus vote has been carried out in order to decide on each proposition. Only 179 queries were associated to several alternative annotations, it means less than 6% of the whole test corpus, with approximately 2 alternatives per query.

The scoring tool developed for the MEDIA project allows the alignment of two semantic representations and their comparison in terms of deletion, insertion, and substitution. It is able to handle alternative representations for each query. The scoring is done on the whole triplet including [`mode` , `attribute name` and `attribute value`]. Different scoring have been performed on the system results. The *Full* scoring used the whole set of attributes, whereas in the *Relax* scoring, the specifiers are no longer considered. Another simplification consists in applying a projection on modes resulting in a mode distinction limited to affirmative and negative (2 modes).

	Full		Relax	
	4 modes	2 modes	4 modes	2 modes
LIA	41.3	36.4	29.8	24.1
LIMSI-1	**29.0**	23.8	**27.0**	21.6
LIMSI-2	30.3	**23.2**	27.2	**19.6**
LORIA	36.3	28.9	32.3	24.6
VALORIA	37.8	30.6	35.1	27.6

Table 2: Results - in terms of understanding error rate - of the MEDIA literal understanding evaluation campaign.

4.2. Results

Table 2 gives the results obtained by the five participant systems in terms of understanding error rates. First it can be observed that the corpus-based training systems obtain better results than the others. The LIMSI-1 system obtains the best performances is based on a totally stochastic model. Concerning the performance of the symbolic systems, a significant part of the errors comes from a bad projection (or translation) into the expected annotation format, and not only from the understanding mistakes (located in conceptual graphs or in logical formula).

The understanding error rates are relatively high : 29% for the best system in *Full* scoring with 4 modes, and 19.6% for the best system in *Relax* scoring with 2 modes. This last result may be compared with the understanding error rate on the ARISE task, with a similar evaluation protocol, which was around 10% on exact transcriptions (Lefevre and Bonneau-Maynard, 2002). The gap in performance between the ARISE and MEDIA tasks may be explained by the number of attributes involved in the models which is much higher for the MEDIA task (83 attributes - 19 specifiers) than for the ARISE task (53 attributes - no specifiers).

The drop in performance between the results obtained with and without the specifiers (*Full vs. Relax*) is very significant for all the systems. As presented in Section 2., the specifiers are used to enrich the flat concept-value representation with hierarchical information. It is worth noting that no significant difference in performance is observed between systems using such a hierarchical representation internally to those obtained with systems implementing a tagging approach. As shown in table 2, the lowest relative increase in error rate (6.8%) is obtained by two systems (VALORIA and LIMSI-1) representing both approaches. Using 4 modes instead of 2 is also a major difficulty for all the systems. The relative increase in error rate imputable to 4 modes ranges from 12% to 23%. This can be partially explained by the fact that signal - which was listened to by the human annotators - is often necessary to disambiguate between interrogative and affirmative mode.

4.3. Error analysis

The attributes on which errors are most frequently done are the reference link attribute (`refLink`). Obviously, the annotation of references represents the most difficult problem on which research teams may have to focus their efforts. This is also true for the connectors identification. Except these two points, the nature of the errors is rather differ-

	Cplx	Rep.	Corr.	Incid.
#occ	136	117	47	26
LIA	54	54	58	59
LIMSI-1	33	38	37	43
LIMSI-2	35	40	41	44
LORIA	47	42	46	47
VALORIA	46	46	53	52

Table 3: Selective understanding error rates in *Full* scoring mode on the subsets of queries containing the main linguistic difficulties: Complex utterances, Repetitions, Corrections and Incidental clauses.

ent among the systems. Therefore, a Rover experiment has been performed in order to seek to exploit the nature of the errors made by the multiple systems and then to reduce the understanding error. The Rover algorithm (Fiscus, 1999) consists in aligning the outputs produced by the different systems in order to produce a graph, and then to select the best scoring attribute at each node. The best ROVER combination achieves a 13% relative improvement in the *Full* mode scoring from the best system results. In the *Relax, 2 modes* scoring mode the relative improvement obtained is more than 17%. In an Oracle mode (ie the system output hypothesis are aligned with the reference sequence), the best ROVER combination obtains a 60% relative improvement from the best system results, resulting in an understanding error rate around 10%.

A meta annotation of the test corpus has been performed by ELDA in terms of linguistic difficulties. Table 3 gives the results for the subsets of queries containing the most significant difficulties in the *Full* scoring mode. *Complex* requests correspond both to multiple requests or requests which are on the borderline of the MEDIA domain. The *Repetition* tag is used when a concept is repeated in the utterance several times with the same value (as in *"the second the second week-end of March"*), whereas *Correction* is used when the concept is repeated with different values (as in *"the second the third week-end of March"*). *Incidental clauses* correspond to sentence portions which temporarily interrupt the current syntactic or meaning sequence of the query (as in *"100 euros as far as I need something confortable 100 euros"*). The understanding error rates become significantly greater for sentences including difficulties. The systems which have got the best results on the whole test set keep the best results for the difficulties. From a the relative point of view, LIMSI-1 and LIMSI-2 systems resist better to complex utterances (less than 17% relative fall) than the other systems (upon 30%). On the other hand with a less than 37% relative fall LORIA and VALORIA symbolic systems are more robust to the incidental clauses than the other systems (upon 43%).

5. Conclusion

The first success of the MEDIA project is that the consortium which involves teams with different speech understanding backgrounds was able to establish a common semantic representation. The 15k user query MEDIA corpus, is fully semantically annotated, with a good quality IAG. Even if a part of the attributes is task-dependent, the representation is generic. Furthermore, it allows to take into account hierarchical relations. Annotation manuals and tools are available and can be reuse for other tasks.

A protocol for speech understanding evaluations has been elaborated by the consortium, allowing an evaluation campaign in June 2005 for literal speech understanding. The corpus also includes the speech signal, so that experiments from speech signal to speech understanding are possible. An evaluation package which includes the corpus along with protocols, scoring tools, and evaluation results will be available and distributed by ELDA.

The MEDIA consortium is currently working on the contextual annotation of the data and the elaboration of a protocol for in-context understanding evaluation.

6. References

H. Bonneau-Maynard and F. Lefevre. 2005. A 2+1-level stochastic understanding model. In *IEEE ASRU*.

H. Bonneau-Maynard et al. 2005. Semantic annotation of the media corpus for spoken dialog. In *ISCA Eurospeech*.

B. Crabbe, B. Gaiffe, and A. Roussanaly. 2003. Une plateforme de conception et d'exploitation de grammaire d'arbres adjoints lexicalises. In *TALN*.

W. Daelemans et al. 2003. Timbl: Tilburg memory based learner, v5.0, reference guide. In *ILK Technical Report ILK-03-10*.

A. Denis, M. Quignard, and G. Pittel. 2006. A deep-parsing approach to natural language understanding in dialogue system: Results of a corpus-based evaluation. In *LREC*.

L. Dybkjaer and N. Ole Bernsen. 1998. The disc approach to spoken language system development and evaluation. In *LREC*.

J.G. Fiscus. 1999. A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (rover). In *IEEE ASRU*.

F. Lefevre and H. Bonneau-Maynard. 2002. Issues in the development of a stochastic speech understanding system. In *ICSLP*.

MADCOW. 1992. Multi-site data collection for a spoken language corpus. In *DARPA Speech and Natural Language Workshop*.

M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1):69–88.

C. Raymond et al. 2006. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48,3-4:288–304.

S. Rosset and D. Tribout. 2005. Multi-level information and automatic dialog acts detection in human-human spoken dialogs. In *ISCA InterSpeech*.

J. Villaneau, J.-Y. Antoine, and O. Ridoux. 2004. Logical approach to natural language understanding in a spoken dialogue system. In *Text, Speech and Dialogue, 7th International Conference*.

M. Walker, R. Passonneau, and J. Boland. 2002. Quantitative and qualitative evaluation of darpa communicator spoken dialog systems. In *ACL/EACL Workshop*.