KRISTINA

A Knowledge-Based Information Agent with Social Competence and Human Interaction Capabilities

H2020-645012

D3.3

Advanced version of the vocal analysis techniques in KRISTINA

Dissemination level:	Public
Contractual date of delivery:	Month 32, 31/10/2017
Actual date of delivery:	Month 33, 03/11/2017
Workpackage:	WP3 Analysis of vocal communication in dialogues
Tasks:	T3.2 (Multilingual Automatic Speech Recognition) T3.3 (Multilingual Text Analysis)
Type:	Report
Approval Status:	Draft
Version:	1.0
Number of pages:	54
Filename:	D3.3AdvancedVersionVocalAnalysis.docx

Abstract

This deliverable – Advanced version of the vocal analysis techniques in KRISTINA – summarizes the research and development work on the vocal analysis technologies done in the context of the KRISTINA project. It focus on the development of multilingual automatic speech recognition and multilingual spoken language understanding systems, which aim, respectively, to transform the user utterances into text, and text into ontological structured representations for processing by the subsequent modules of the KRISTINA platform.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	10.10.2017	First draft	Thiago Fraga da Silva
0.2	18.10.2017	Second draft	Thiago Fraga da Silva
0.3	24.10.2017	Integration of partner contributions	Thiago Fraga da Silva
0.4	30.10.2017	Internal Review	Florian Lingenfelser
0.5	31.10.2017	Revision of the draft according to the comments of the Internal Reviewer	Bianca Vieru, Mireia Farrús, Simon Mille, Alicia Burga, Stamatia Dasiopoulou, Alp Öktem, Mónica Domínguez
0.6	02.11.2017	Final draft	Bianca Vieru
1.0	03.11.2017	Revision of the final draft and upload to the Portal	Leo Wanner

Author list

Organization	Name	Contact Information
VR	Bianca Vieru	vieru@vocapia.com
VR	Thiago Fraga da Silva	thfraga@vocapia.com
VR	Yvan Josse	josse@vocapia.com
VR	Nidia Hernandez	nidia@vocapia.com
VR	Julien Despres	despres@vocapia.com
VR	Viet-Bac Le	levb@vocapia.com
VR	Lori Lamel	lamel@vocapia.com
VR	Abdel Messaoudi	abdel@vocapia.com
VR	Jean-Luc Gauvain	gauvain@vocapia.com
UPF	Simon Mille	simon.mille@upf.edu
UPF	Stamatia Dasiopoulou	stamatia.dasiopoulou@upf.edu
UPF	Mireia Farrús	mireia.farrus@upf.edu
UPF	Alicia Burga	alicia.burga@upf.edu
UPF	Leo Wanner	leo.wanner@upf.edu
UPF	Mónica Domínguez	monica.dominguez@upf.edu
UPF	Alp Öktem	alp.oktem@upf.edu
CERTH	Georgios Meditskos	gmeditsk@iti.gr
semFYC	Chiara Baudracco	cbaudracco@semfyc.ces
EKUT	Valérie Sarholz	valerie-charlotte.sarholz@med.uni-tuebingen.de

Executive Summary

This document describes and illustrates the functionality of the vocal analysis techniques as developed during the project KRISTINA. The audio data collected by the user partners throughout various recording sessions is described. A summary of the work done for adapting the automatic speech recognition and spoken language understanding systems to the KRISTINA task is provided. The research and development described here contributed to significant gains into the speech recognition and language analysis modules towards the development of the final KRISTINA system. This document contributes to the achievement of MS5.

Abbreviations and Acronyms

ASR	Automatic speech recognition
DRK	Deutsches Rotes Kreuz, Kreisverband Tübingen (project partner)
EKUT	Eberhard Karls Universität Tübingen (project partner)
LAS	Labeled attachment score
OOV	Out of vocabulary
OWL	Web ontology language
RDF	Resource description framework
semFYC	Sociedad Española de Medicina de Familia y Comunitaria (project partner)
SLU	Spoken language understanding
UC	Use case
UPF	Universitat Pompeu Fabra (project partner)
VR	Vocapia Research (project partner)
WER	Word error rate

Table of Contents

1 INTRODUCTION

This document describes the advances achieved in tasks T3.2 (*Multilingual Automatic Speech Recognition*) and T3.3 (*Multilingual Text Analysis*) in the context of the KRISTINA project. The project aims at creating a multimodal communication interface between a virtual agent and a human user. This interface is supported by several core technologies, including: automatic speech recognition (ASR), spoken language understanding (SLU), gesture analysis, emotion recognition, multimodal fusion, information retrieval, dialogue management, avatar communication generation, natural language generation, and text-to-speech synthesis.

The focus of this deliverable is on the vocal analysis technologies, which aim at converting human spoken utterances from speech to text (ASR), and from text to ontological representations (SLU) that can be properly handled by the dialogue manager. The project covers spoken dialogues in the basic and health care domain in five languages: Arabic, German, Polish, Spanish and Turkish. The research and development work performed during the project is presented here. The advances contributed towards the implementation of the first and second prototypes, and the final system.

This document is structured as follows. Section 2 summarizes the available corpora used to drive research and development of ASR and SLU technologies. Section 3 describes the development strategies applied in adapting the ASR systems to the KRISTINA task. Section 4 reports on the advances made on SLU technologies, with focus on the development of multilingual dependency parsers. Section 5 presents the work planned towards the implementation of the final demonstrator. Conclusions are provided in Section 6.

2 CORPUS DESCRIPTION AND ANNOTATION

Most of the vocal analysis technologies presented in this work are based upon statistical methods, which generally rely on large amounts of annotated data to estimate reliable model parameters. As far as possible, the data should be representative of the target task, which is linked to the language, topic and domain, speaking style, application, and audio quality. The vocal analysis modules aim at processing user requests and statements on the basic and health care domain in the following languages: Arabic, Spanish, German, Polish, Spanish and Turkish. Throughout the project, KRISTINA's user partners, DRK, EKUT and semFYC, have recorded spoken dialogues in these five languages. Such recordings were manually and semi-automatically annotated to drive the development of automatic speech recognition (ASR) and spoken language understanding (SLU) modules.

In addition to the manual transcriptions produced from the audio recordings, other textual data were collected in the context of KRISTINA. Written sources consist of manually produced dialogues and data on different health topics were automatically retrieved from the Web. The following sections summarize the characteristics of the KRISTINA specific corpora as well as the annotation procedures adopted in the project.

For language modeling, broadcast and conversational speech transcriptions, as well as other Web texts available prior to KRISTINA were used. Such non-related KRISTINA data constitute around 99% of the training data on each language.

2.1 Acoustic corpora

KRISTINA's user partners organized several recording sessions between months M3 and M30. The recordings took place in Spain (Arabic and Spanish) and Germany (German, Polish and Turkish).

The dialogues are in line with the project use cases, which cover various topics of the basic and health care domain. The use cases were defined in Work Package 8, and vary according to the language, the system version (prototypes or final system), and the role of the virtual agent in the conversation (e.g. social companion, health expert, mediator, nursing assistant, receptionist). The degree of spontaneity varies across the recording sessions. In some cases, the participants were encouraged to freely discuss about a given topic. In other recordings, the participants should derive their conversation from the use case dialogue flows. Finally, few recordings were obtained from pre-defined utterances. The participants were oriented to avoid overlapping speech. Each dialogue was performed between two native speakers, which might have been influenced by the language of their hosting country (the case of Arabic, Polish and Turkish speakers). The influence of such phenomenon is not evaluated in this work. Technical characteristics of recordings also changed during the project. Different types of microphones were used; recording via the SSI[1] framework and other hardware-specific tools was assessed; the characteristics of the room in which recordings took place also changed to reduce reverberation and background noises. In summary, the overall quality of recordings evolved during the project lifetime.

[1] The social signal interpretation framework (https://hcm-lab.de/projects/ssi/)

Table 1 presents statistics of the audio recordings collected within the project. As much as possible, KRISTINA data was used for ASR and SLU system development and testing. As these audio recordings were naturally not available at the beginning of the project, a significant part of ASR system development and evaluation was realized using broadcast speech and conversational telephone speech data sets, available prior to KRISTINA. The share of the external data in the training datasets varies from language to language. For instance, for Spanish, more than 400 hours, for German about 150 hours and for Turkish about 100 hours of transcribed external audio data were used.

Language	Session	Number of dialogues	Number of speakers	Audio duration (min)
Arabic	M3	18	8	46
	M17	16	6	39
	M18	13	5	63
	M23	17	4	70
	M24	52	6	122
	M28	18	3	56
	M30	7	3	30
	Total	***141***	***9***	***426***
German	M3	14	8	70
	M10	32	7	125
	M13	33	5	122
	M16	54	11	150
	M17	21	3	49
	M18	20	4	52
	M20	14	3	59
	M21	18	3	50
	M25	19	4	55
	Total	***225***	***28***	***722***
Polish	M3	2	2	16
	M12	24	4	86
	M13	45	4	105
	M17	35	6	105
	M24	30	4	124
	M27	21	3	112

	M30	30	5	151
	Total	***187***	***16***	***699***

Table 1. Audio corpus recorded and transcribed in KRISTINA. The audio duration is obtained after manual data selection (continue on the next page).

Language	Session	Number of dialogues	Number of speakers	Audio duration (min)
Spanish	M3	9	8	17
	M9	28	9	169
	M10	1	18	37
	M12	8	2	32
	M14	16	3	70
	M17	81	13	279
	M24	2	2	14
	M27	21	4	43
	Total	***166***	***45***	***661***
Turkish	M3	12	9	78
	M12	9	2	43
	M16	18	4	42
	M23	26	3	96
	M24	44	5	87
	M25	10	4	34
	M26	18	3	61
	M27	22	4	48
	M28	27	4	91
	M30	48	6	162
	Total	***234***	***19***	***742***

Table 1. Audio corpus recorded and transcribed in KRISTINA. The audio duration is obtained after manual data selection (continuation).

2.2 Acoustic data annotation

The acoustic data collected in KRISTINA was manually annotated. Native speakers transcribed the audio data following project-specific guidelines. They include the use of the standard orthographical form of a word (even when mispronounced), the use of specific symbols to

represent hesitations and truncated words, among others. Transcriptions were validated to ensure the guidelines were applied.

The KRISTINA data are comprised of dialogues between two persons, one having the role of the "user" and the other of the "agent". A subset of the Spanish data was manually annotated with speaker turns and the appropriate speaker labels in order to estimate how much of the data represents user utterances, as it consists of the actual target data of the ASR and SLU modules in KRISTINA. Table 2 shows the specific user and agent information for the Spanish data recorded at month M9. About a quarter of the data consists of the user side of conversations on this particular data set. Kristina users interact only by short questions, while agents give long and complete answers.

Speaker	Audio duration (min (%))	#words (x 1000 (%))	#different words (x 1000 (%))
Agent	123 (73%)	20.6 (76%)	2.6 (84%)
User	46 (27%)	6.7 (24%)	1.2 (39%)
Total	***169 (100%)***	***27.4 (100%)***	***3.1 (100%)***

Table 2. Spanish M9 data set split between agent (KRISTINA) and user.

The punctuation generation module is trained with both lexical and acoustic features. The neural network model predicts punctuations between each word looking at the word sequence and also their acoustic properties. To be able to train such a model, a prosodically annotated speech corpus is needed. Word alignments are necessary to assign acoustic properties to each word. The model is trained only with spoken data.

Spanish has been chosen as the initial language to employ the punctuation generation module. As speech data source, the Glissando corpus (Garrido, 2013) was used. The "task-oriented" section of the corpus consists of high-quality recordings of dialogues between two speakers: "asker" and "giver", towards a goal. The "asker" requests some information and the "giver" replies appropriately. This corpus has 12 speaker pairs, each pair talking approximately 10 minutes about three different topics: university, transportation and tourism. The "informal dialogue" section of the corpus is less oriented. Six speaker pairs are recorded having casual conversations. Each recording has between 5 and 16 minutes. In total 426 minutes of conversation were used to train the punctuation model. Word time information was obtained via forced alignment. Acoustic-prosodic features were obtained using the methodology described in Farrús (2016).

The words from the KRISTINA corpus that are not covered in the Glissando corpus were added in order to adapt the punctuation model to the KRISTINA domain. In the second phase of this study, the KRISTINA data was used to train another punctuation model, which was compared to the model obtained with the Glissando corpus.

2.3 Textual corpora

Simulated dialogues tailored to the first prototype use cases were manually created via a dedicated Web-based platform accessible to KRISTINA partners. Table 3 shows the statistics of the language specific corpora.

Although the manually produced data (audio transcriptions and simulated dialogues) are the best match to the KRISTINA application, the amount of data is relatively small. Additional data was automatically retrieved from Websites having health and basic care related content. These sources were inventoried and validated by native speakers collaborating in the consortium. Additional textual corpora available prior to the project were also used for system development. It includes manual transcriptions of broadcast speech and conversational speech and Web-based textual sources. The available data was cleaned, normalized and used for adapting the language models of the different ASR systems. More information about the text corpora used for language model estimation is provided under the ASR system descriptions for each language in Section 3.

Language	Scenario	#words	#unique words
Arabic	3	860	513
	4	11.8k	3.4k
German	1	2.4k	994
	2	3.0k	1.1k
Polish	2	3.6k	1.6k
Spanish	3	2.1k	914
	4	1.7k	733
Turkish	1	540	317

Table 3. Written simulated dialogues collected from KRISTINA partners. For each language and scenario, the number of words and unique words obtained prior to text normalization is indicated. Unit k=x10^3.

2.4 Textual corpora annotation

The textual output provided by ASR is analyzed by the SLU systems in order to transform the natural language representation of the user utterances into structured ontological representations, so that appropriate system responses can be subsequently inferred. Along the project, multilingual corpus annotation has been done, for training the required multilingual dependency parsers for SLU processing.

The development of both statistical analyzers and generators is based on the Meaning-Text Theory (MTT) (Mel'čuk, 1988), a linguistic framework that foresees various levels of representation, all based on dependencies. Having good quality data at hand is crucial for training efficient machine-learning systems, which is why Task 3.3 (*Multilingual Text Analysis*) is partly centered on the annotation of resources.

Parallel annotation of morphological structures (MorphS), surface-syntactic structures (SSyntS), and deep-syntactic structures (DSyntS) was carried out in the different languages of KRISTINA. MorphS consists of a linearized structure, each node with all the morphological information needed to be inflected (*Figure 1*); SSyntS consists of a non-linearized structure containing all the words in a sentence (even the functional ones) related through language-specific relations (*Figure 2*); DSyntS consists of a structure in which just the lexemes (meaningful units) appear, related through language-independent relations, which

distinguish arguments (I, II, III…) from modifiers (ATTR, APPEND) and coordinations (COORD) (*Figure 3*). This level needs to be revised manually for having parallel SSynt and DSynt structures, to train a statistical parser. Regarding the KRISTINA data specifically, the analysis grammars need to check that this level is right.

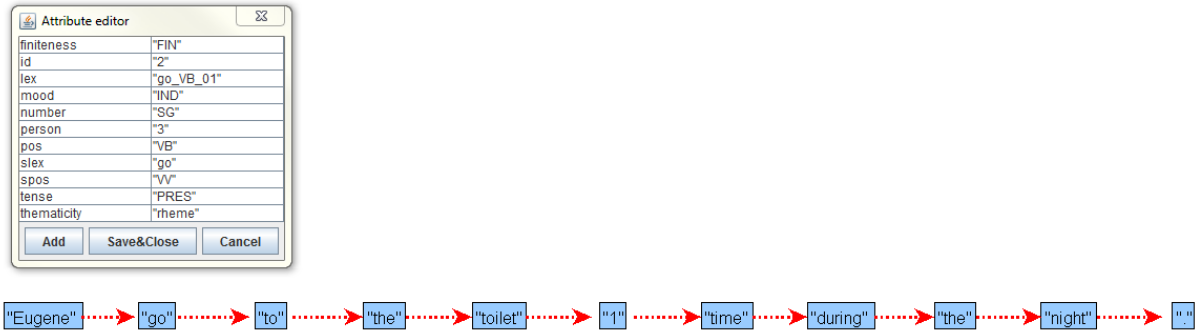

Figure 1: A linearized morphological structure that corresponds to the sentence "Eugene goes to the toilet one time during the night".

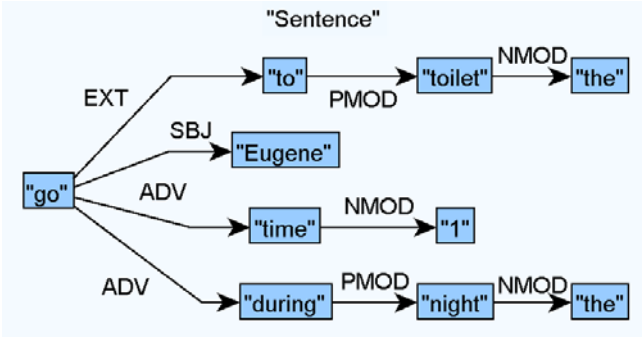

Figure 2: A surface-syntactic structure that corresponds to the morphological structure in Figure 1.

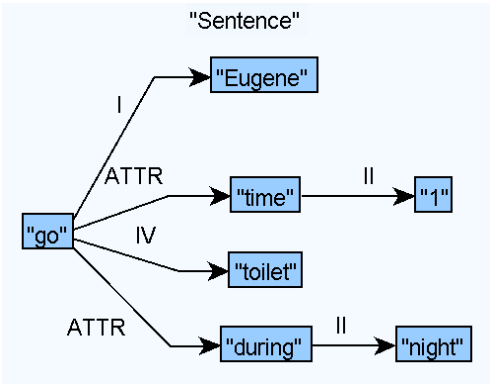

Figure 3: A deep-syntactic structure that corresponds to the surface-syntactic structure in Figure 2.

All the layers and the original texts need to be aligned sentence by sentence and node by node, which can be done thanks to unique identifiers associated to each sentence and node. The morphologic annotation (see the attribute editor in Figure 1) consists of features associated to each word of the sentence, including coarse-grained and fine-grained part-of-speech, gender, number, tense, aspect, finiteness, mood, person, etc. On this layer only, the

order of the words is kept. The surface-syntactic annotation (Figure 2) consists of dependency trees with all the words of a sentence linked by idiosyncratic, therefore language-dependent, relations. At the deep-syntactic layer (Figure 3), all functional (i.e. non-meaningful) units are removed: definite and indefinite determiners and auxiliaries are replaced by attribute/value pairs on the concerned nodes, while punctuations and governed prepositions, and conjunctions are simply removed. In addition to attributive and coordinative relations, the dependency relations also encode predicate-argument information, through the assignment of an argument slot in the valency (sub-categorization *framework*) of its governor predicate.

We currently have available multi-layered corpora for Spanish and English at different stages of advancement; German, Polish and Turkish have also been tackled. Arabic has not been annotated, due to difficulties with pre-processing resources (morphological tagger and parser).

For English, we use the Penn Treebank 3 (Marcus et al., 1994) converted to dependency trees (Johansson & Nugues, 2007), which we use as surface-syntactic and morphologic annotations. We automatically derived a first version of the deep annotation from the surface annotation using graph transduction grammars implemented in the MATE environment (developed by TALN-UPF). During the mapping, we removed all determiners, auxiliaries, "*that*" complementizers, infinitive markers "*to*", punctuations and functional prepositions. The treebank currently contains 41,678 sentences (1,015,843 tokens at the surface-syntactic and morphologic layers, 768.865 tokens at the deep-syntactic layer). The annotation is currently being reviewed and is continously being improved.

For German and Polish, we follow the same method as for English; we start with the existing surface-syntactic corpora (respectively the TIGER corpus (Brants et al., 2002), and the Polish Dependency Bank (Wolinski et al., 2011)) from which we automatically derive the deep-syntactic annotation thanks to graph transduction grammars. Since the surface-syntactic annotation is quite different from the English one, it is not possible to re-use the English mapping: the dependency tag set and the annotation scheme are not the same. Given that the SSynt relations established in the training corpora are well defined, just the output of the transduction has been manually reviewed by native speakers. We have approximately 6,500 manually reviewed DSyntSs in Polish and 6,000 in German.

Regarding Turkish, we have used as base the METU-Sabancı treebank (Atalay et al., 2003). However, given that the SSynt tagset is not well defined, it has been necessary to develop a new SSynt tagset and its corresponding annotation guidelines, before a manual revision could be applied. Thus, during the project 2,500 SSynt and 2,500 DSynt structures have been manually reviewed.

For Spanish, we are extending the AnCora-UPF dependency Treebank (Mille et al., 2013)[2]. In the AnCora-UPF, each layer is independently annotated, and the annotation is manually validated. For the surface-syntactic annotation, there are several dependency tag sets available, with many or few tags. The different dependency tag-granularities obtained different annotator-agreement values, but they were between 89.4% and 92.26% from the most fine-grained to the most coarse-grained. Those different granularities will be tested in

[2] The corpus underlying this treebank is the 2008 version of AnCora (Taulé et al., 2008).

order to find the optimal tag set for the statistical generation. The treebank currently contains around 5,000 sentences (131,126 tokens at the surface-syntactic and morphologic layers, 89,072 tokens at the deep-syntactic layer), and we aim at reaching 6,500 sentences by the end of the project. In the framework of KRISTINA, we manually annotated about 700 sentences coming from dialogues in the medical domain in order to add them to the training set and improve the results of our parser on spoken language (see Section 4.4).

Table 4 summarizes the total number of sentences comprising the annotated corpora compiled within KRISTINA; in parallel, the annotation of Arabic corpora has already started. It is worth noting that there is no KRISTINA data included, but this was done in the context of KRISTINA (for having the possibility of training statistical DSYnt parsers).

Language	Representation Layer	Number of Sentences
German	DeepSynt	6000
Polish	DeepSynt	6500
Spanish	SurfaceSynt, DeepSynt	5000
Turkish	SurfaceSynt, DeepSynt	2500

Table 4. Summary of annotated corpora compiled within KRISTINA.

3 AUTOMATIC SPEECH RECOGNITION

This section briefly describes the baseline ASR systems and summarizes the work performed to build systems adapted to KRISTINA. Once the task specific audio and textual data are available and have been prepared, adapting a system to a task is usually comprised of: tuning the model parameters to the target data; tuning the recognizer parameters; building specific pronunciation dictionaries; assessing different model and system topologies; developing a specific post-processing module; performing iterative system evaluation. In addition, developing a real-time dialogue system requires specific development strategies as described for instance in Gauvain et al. (1996) and Rosset et al. (1999).

3.1 Baseline systems

State-of-the-art ASR systems are based on statistical learning methods and have five main components: an audio partitioner, an acoustic model, a statistical language model (LM), a pronunciation dictionary, and a word recognizer. More specifically, the audio partitioner generates a sequence of non-overlapping segments and groups them into clusters, each representing a speaker in a given acoustic condition (Gauvain et al., 2001). The acoustic model provides a likelihood function of the signal given a sequence of phoneme units. The language model attempts to capture regularities in the underlying process of natural language generation, providing a prior probability for a given word sequence. The pronunciation dictionary links the acoustic-level representation with the word-sequence output. The word recognizer provides the most likely sequence of words provided an input acoustic signal and the acoustic, language and pronunciation models.

In this work, a broadcast audio partitioner was used during system development. The KRISTINA prototypes do not use an audio partitioning, given the user segments the audio via a push-to-talk button. Automatic voice activity detection has been assessed in the context of Work Packages 4 and 7. The following sections provide the main characteristics of the baseline models.

3.1.1 Acoustic models

As described in (Gauvain et al., 2002), the acoustic models are word position dependent triphone-based left-to-right tri-state Hidden Markov Models (HMM), with output observation densities given by Gaussian Mixture Models or (Deep) Neural Networks (DNN). The acoustic training corpora comprises audio recordings and their associated transcriptions.

The input feature vector of the GMM/HMM acoustic models is based on the Perceptual Linear Predictive (PLP) (Hermansky, 1990) features concatenated with Multilayer Perceptron bottleneck (Grézl et al., 2007) features. The input feature vector of the DNN/HMM models consists of PLP feature vectors. About 10k HMM tied-states are used for the acoustic models (GMM and DNN-based). For both types of models, the acoustic feature vector is obtained after mean and variance normalization. Speaker adaptive training (Gales, 2001) is also applied.

3.1.2 Language models

The role of the language model is to capture regularities in the underlying process of natural language generation, providing a prior knowledge of the language. More specifically, statistical language models provide a prior probability of a given sequence of words. In this work, back-off *n*-gram language models are used. They are estimated in two steps: frequency counting of word sequences (*n*-grams), followed by probability smoothing to improve model generalization (Kneser and Ney, 1995). Language model training data includes manual transcriptions of recordings, simulated dialogues and written sources gathered from the Web.

Specific statistical LMs are generally estimated for each application, given that the word sequence probabilities are dependent on the language, the domain (e.g. healthcare, economics, politics), the task (e.g. dialogues, news, meetings) and the style of speech (e.g. prepared, conversational) (Rosenfeld, 2000). The language models are obtained as follows. First, a component LM is built for each available source individually (transcriptions, dialogues, news, blogs, *tweets,* Web sites). These component LMs are then combined via linear interpolation, with coefficients automatically estimated in order to maximize the likelihood of the development (held-out) data. This development data generally consists of manual transcriptions of audio recordings matching the target application.

3.1.3 Word recognition

Word recognition is performed in a single pass, that is, the audio data is processed just once. This step uses the acoustic model, the pronunciation dictionary and a basic language model. It generates a word graph, which is further re-processed with larger language models. The final raw transcriptions are obtained via consensus decoding (Mangu et al., 1999). A post-processing module is used to add punctuation and capitalize utterances. Post-processing is language dependent and makes use of specialized models.

Besides the transcription and punctuation, the ASR output includes multi-level metadata information such as: time markers for speech segments and words; confidence scores for recognized words and the identified language; overall confidence scores; speaker turns and speaker labels. The confidence measures are numbers between 0 (not confident at all) and 1 (highly confident) and can be useful to help on the decisions of the subsequent modules, namely language analysis.

3.1.4 System evaluation

The quality of ASR systems is generally measured in terms of the word error rate (WER). After automatically aligning the system output (hypothesis) with the manual transcriptions (reference), three types of recognition errors are computed: substitutions (S), insertions (I) and deletions (D). These errors are used to calculate the WER, as follows:

$$WER = \frac{S + D + I}{N}$$

where *N* is the number of words in the reference.

Additional measures can be used to assess the recognition vocabulary and the language model. The out-of-vocabulary rate (OOV) gives the average ratio of words in the

development set that are unknown and consequently cannot be recognized by the system. The perplexity gives the normalized cross-entropy of the development data with respect to a language model, providing a broad idea of how difficult the data is for a language model.

For all three measures, WER, OOV and perplexity, the lower the better.

3.2 Real-time processing

Vocapia's ASR broadcast off-the-shelf systems can operate in two modes. On the *standard* mode, an audio file is given to the system input and the output transcriptions are returned once recognition ends. Processing time takes roughly the signal duration. On the *streaming* mode, the system receives an input audio stream and returns transcriptions on another stream as soon as they become available, with a latency varying from 10 to 20 seconds for

Both modes, standard and streaming, use the same type of acoustic models. To extract the input vectors of such models, feature normalization and unsupervised speaker adaptation are required. They are both applied within segments of speech, which need to be fully read before word recognition starts. Generally speaking, the longer the segment is, the more effective normalization and adaptation are. Neither of the modes is well adapted for dialogue applications, such as in KRISTINA, in which getting a fast and accurate response is essential from a user's point of view.

The solution considered in KRISTINA was to build acoustic models that do not rely on segment-wise operations. The main advantage of this setup is that word recognition can effectively start as soon as the very first speech frame vectors are available at the input of the system. In addition, there is more flexibility for adjusting the system parameters and finding a trade-off between recognition accuracy and latency.

A series of experiments was realized to adapt the Spanish system to a real-time dialogue application. The best system setups were then applied to the other four languages. The first experiments were performed to assess the impact of dropping out pitch features, feature normalization and unsupervised adaptation. Experiments with improved acoustic models are described afterwards. An evaluation of processing time and latency is summarized at the end of this section.

3.2.1 Assessing the influence of segment-based processing steps

Four DNN/HMM based acoustic models were built for Spanish using slightly different input features. They are all based on PLP analysis, but differ on whether or not speaker adaptation, pitch features or feature normalization are used. Table 5 summarizes the recognition results obtained on broadcast data. This comparison was done using DNN models obtained via cross-entropy training.

Removing speaker-based adaptation (PLP-1) leads to an absolute WER loss of 0.4% compared to the baseline (PLP-0). Further removal of pitch features (PLP-2) and cluster-based feature normalization (PLP-3) lead to additional losses of 0.1% and 0.4% in WER. In summary, the combined use of these three techniques contribute for a performance difference of about 0.9% absolute (7.5% relative) on Spanish broadcast data (cf. PLP-0 and PLP-3).

System	Speaker adaptation	Pitch	Feature normalization	WER

PLP-0 (baseline)	Yes	Yes	Yes	11.1
PLP-1	-	Yes	Yes	11.5
PLP-2	-	-	Yes	11.6
PLP-3	-	-	-	12.0

Table 5. WER (%) of acoustic models trained using different PLP-based input feature vectors on broadcast Spanish data.

3.2.2 Temporal Pattern (TRAP) based acoustic models

The baseline acoustic models are based on PLP features. As an alternative, temporal pattern (TRAP) spectrum analysis (Schwarz et al., 2004) can be used. Typically, TRAP features are used as input of multilayer Perceptron neural networks, from which discriminatively trained feature vectors can be extracted from one of the hidden layers (the *bottleneck*) (Grézl et al., 2007; Fousek et al., 2008a). The resulting TRAP-bottleneck features can then be used as input for GMM/HMM or DNN/HMM acoustic models. TRAP-bottleneck features have been successfully used in many ASR tasks, and have been proven especially useful when combined with short-term features (Fousek et al., 2008b).

Two TRAP bottleneck models, with (TRAP-1) and without (TRAP-2) segment-wise feature normalization, were built for Spanish. These models were assessed on broadcast data, and the recognition results are summarized in Table 6. In both cases, the TRAP models obtained better recognition performance than their PLP counterparts. The performance gap between the baseline and TRAP-2 is 0.5% absolute, even though the input vector of TRAP-2 does not use pitch features, feature normalization and speaker adaptation.

System	Speaker adaptation	Pitch	Feature normalization	WER (cross-entropy)	WER (sMBR)
PLP-0 (baseline)	Yes	Yes	Yes	11.1	10.3
PLP-2	-	-	Yes	11.6	-
PLP-3	-	-	-	12.0	-
TRAP-1	-	-	Yes	11.0	10.2
TRAP-2	-	-	-	11.6	10.9

Table 6. WER (%) comparison of PLP and TRAP based cross-entropy and sMBR acoustic models on Spanish broadcast data.

The TRAP models were further refined via discriminative sMBR DNN training (Vesely et. al, 2013). Results are shown in the last column of Table 6. The performance gap between the baseline and TRAP-2 sMBR models remains around 0.6% absolute.

3.2.3 Relative spectral (RASTA) filtering

Relative spectral (RASTA) filtering was introduced by Hermansky and Morgan (1994) as a means to reduce speech signal distortion due to environmental noises. Assuming that the communication channel noise is steadier than speech, it makes use of a linear band-pass filter to deconvolve speech and noise. RASTA filtering was shown to be especially useful in

mismatching training and testing conditions. As observed by Hermansky and Morgan, RASTA filtering is somehow related to cepstral mean subtraction in the sense that both are able to remove a fixed bias in the cepstral domain.

RASTA filtering was assessed on the Spanish broadcast recognition task. The recognition performance of the TRAP-RASTA model is shown in Table 7. The TRAP-RASTA model obtained a WER of 10.5%, that is 0.2% behind the PLP-0 baseline.

System	Speaker adaptation	Pitch	Feature equalization	WER
PLP-0 (baseline)	Yes	Yes	Normalization	10.3
TRAP-1	-	-	Normalization	10.2
TRAP-2	-	-	-	10.9
TRAP-RASTA	-	-	RASTA	10.5

Table 7. WER (%) comparison of the baseline, TRAP and TRAP-RASTA acoustic models on broadcast Spanish data.

3.2.4 Evaluation on KRISTINA data

The different acoustic modeling techniques described in the previous sections were also evaluated on the KRISTINA Spanish M9 data set. A language model adapted to the KRISTINA domain was used here. Table 8 summarizes the results obtained with sMBR DNN/HMM acoustic models.

System	Speaker adaptation	Pitch	Feature equalization	WER (user,agent)
PLP-0 (baseline)	Yes	Yes	Normalization	18.8 (27.4, 15.9)
PLP-3	-	-	-	25.1 (38.4, 20.7)
TRAP-2	-	-	-	25.0 (37.5, 20.8)
TRAP-RASTA	-	-	RASTA	20.8 (31.1, 17.2)

Table 8. WER (%) comparison of different acoustic models on KRISTINA M9 data set. For each system, the overall WER is shown, as well as the WER on the user and the agent parts.

As for the broadcast task, the baseline model obtains the best overall recognition performances on KRISTINA data. This could be expected, given that the input vector of the baseline model includes pitch features and is refined via feature normalization and speaker adaptation. The performance gaps between the baseline and the models designed for real-time processing (PLP-3, TRAP-2, TRAP-RASTA) are higher for the KRISTINA data than observed for the broadcast data. For instance, the WER relative difference between TRAP-2 and the baseline is 25% on KRISTINA, but only 6% on broadcast data. This is probably due to a mismatch between training and testing conditions. The Spanish acoustic models are trained on a few hundreds of broadcast data, which do fully match the KRISTINA data in terms of environment conditions and speaking style. The difference between the TRAP-RASTA model and the baseline is about 7% relative (from 18.8% to 20.8%).

The WER performance on the user side of the conversation is almost twice as high as the performance on the agent. The agent, which has a role of a doctor in the M9 Spanish data set, employs a more formal speaking style and generally speaks more clearly than the user. In other words, the agent type of data is closer to the broadcast data used in acoustic model training. In addition, the volume of agent data is greater than the volume of user data (cf. Table 2). Acoustic model adaptation is required to improve recognition performance on the user data, the main target of ASR in the context of the project. Unfortunately, the amount of acoustic data available for adaptation is relatively small compared to the amount of broadcast data used in acoustic modeling. An alternative is to use advanced techniques to increase the model robustness to the changes in environment conditions, such as acoustic data augmentation (Lippmann et al., 1987; Deng et al., 2000). Data augmentation experiments are described in Section 3.3 .

3.2.5 Word recognizer and processing time

The acoustic modeling techniques described in the previous sections were evaluated using the baseline word recognizer. The word recognizer was also adapted to KRISTINA. First, the recognizer was modified to process an input stream to start decoding once the first acoustic frames become available. Recognition was also simplified. First, a 2-gram language model is used to generate a word graph, which is re-scored with a 3-gram model. Consensus decoding is applied to obtain the final raw transcriptions. Pronunciation and language models are kept in memory to speed-up processing. Post-processing is applied to add punctuation.

These changes enable recognition to run faster and to minimize latency, but also affect the recognition accuracy. Figure 4 shows the WER on the Spanish M9 KRISTINA data as a function of decoding time, represented here by the real-time factor (RT), which is roughly the processing time normalized by the duration of the test data. This graph was obtained by varying several decoding parameters.

The reference for these experiments is the system TRAP-RASTA shown in Table 8. Such a system uses three language models, runs at 2.0xRT on a single core and has a WER of 20.2%. The use of two language models increases the WER to about 21.6% for a 2.0xRT system. The decoding parameters were tuned to enable real-time recognition (processing at 1.0xRT or lower), leading to a WER of 22.9%. Further development improvements contributed to reduce the WER to 22.1%. In summary, there is an absolute WER difference of about 1.9% (from 20.2% to 22.1%) after adapting the recognizer for real-time processing.

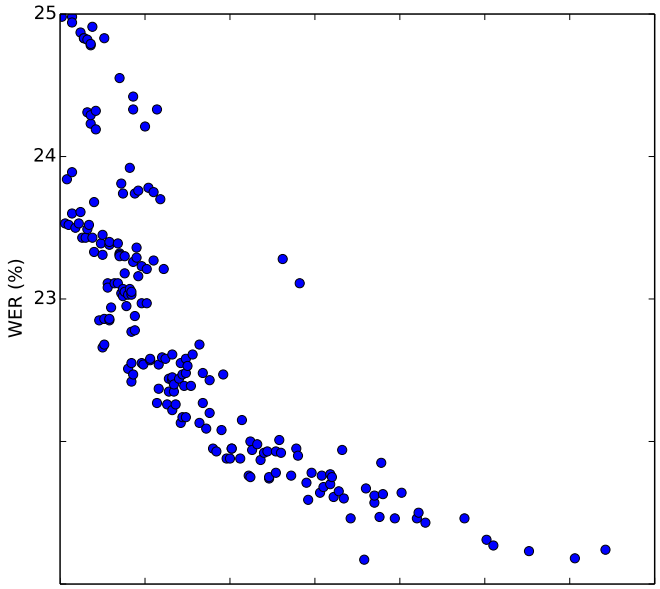

Figure 4: WER on the M9 Spanish data as a function of decoding time.

Additional tests were performed to measure latency. A local server was used to capture the M9 KRISTINA data via streaming. Time was measured between the different steps, namely word recognition, consensus decoding and post-processing. As processing time and latency depend on the test data, statistics were measured for each utterance. For a total of 1157 utterances, the average global recognition latency is 0.441s, and the median 0.40s. Post-processing alone collaborates with almost half of latency, or 0.198s in average. In these experiments, post-processing uses a lexical-based approach, in which the probability of having a punctuation mark is obtained from an n-gram language model. Section 3.4 describes the research work on the use of lexical and acoustic features and neural network models for punctuation determination. Such a system is not yet integrated into the KRISTINA platform.

3.3 Acoustic data augmentation

The acoustic models used in KRISTINA are trained on several hundreds of hours of transcribed speech data in each language. The data consists of broadcast conversations and telephone speech, covering a variety of speakers, accents and topics. Although the large volume, the acoustic training data is distant at some extent from the KRISTINA domain. For instance, the KRISTINA user requests are shorter utterances than what is usually found in broadcast shows. The quality of KRISTINA recordings is also lower than the majority of the training data due to the lower quality of material, presence of background noise and reverberation. In addition, the acoustic training data contain speech produced for human-to-human communication (newswires, dialogues, interviews, conversations, etc), while in KRISTINA the user speaks to a virtual agent. Speaking to a machine may influence the speaking style to a less natural and less spontaneous interaction, as reported into the

subjective evaluation of the first prototype. The influence on speech recognition accuracy was not studied into this work though.

To cope with the mismatching between training and test conditions, acoustic data augmentation was assessed (Ragni et al., 2014; Cui et al., 2015). The goal of such technologies is to increase model generalization and, subsequently, the system robustness to unseen conditions. First, artificial copies of the training data are generated by varying the signal characteristics, (such as pitch, volume or speech rate), or adding background noises. Then, the modified copies are used along with the original data into model estimation.

Acoustic data augmentation was assessed on the KRISTINA Turkish task. The acoustic copies were generated by varying the speech rate, the volume and reverberating the input signal. A single copy of the training data was generated. Therefore, the amount of training data doubled. Several training setups were assessed to take into account the extended training set. The recognizer was optimized to minimize the word error rate on the KRISTINA development data.

The best model obtained a significant gain into both broadcast and KRISTINA tasks. On the KRISTINA development set, the WER of the real-time system reduced from 38.9% to 33.7% with the improved acoustic models.

3.4 Punctuation determination

The punctuation module currently integrated in KRISTINA is based on lexical features. Probability of having punctuation marks or not is derived from a specific punctuation language model. This section describes the research done as an attempt to automatically punctuate the raw transcriptions produced by the ASR system using neural networks and lexical and acoustic features. Later it draws a schema how it will be incorporated into the KRISTINA pipeline.

3.4.1 Neural network based punctuation model

Raw automatic speech recognition (ASR) output lacks of punctuation which is essential for readability, and in the case of a number of tasks, subsequent automatic processing. Punctuation of spoken texts is influenced by two intertwined linguistic phenomena: (1) syntax and (2) prosody. Syntax determines the distribution of punctuation marks in accordance with the grammar of a language. Prosody realization in speech (for instance word grouping, pausing, emphasis, rising-falling intonation, etc.) tends also to signal the position and type of punctuation marks.

Approaches to generate punctuation in transcribed speech are driven by either syntactic or prosodic criteria. A recent work (Tilk, 2016) takes into account textual features together with pause durations between words. However, like other state-of-the-art models, it ignores other prosodic phenomena such as fundamental frequency (f0) and intensity. Also, their two-stage approach can create a bias towards written data instead of spoken data, notably because the amount of written data is generally considerably greater than the amount of spoken data available for model estimation.

We have developed a methodology that is able to take lexical and prosodic information in parallel for punctuation generation in transcribed speech (Öktem, 2017). In the proposed configuration, the model can be fully trained on spoken data. Also, this way we can integrate

any desired feature (lexical, syntactic or prosodic) and test their influence in punctuation placement.

The overview of the proposed neural network model is given in *Figure 5*. It consists of gated recurrent unit (GRU) layers (Cho, 2014) that process the words in both directions and prosodic features in the forward direction. Prosodic features are normalized and discretized before represented in the embedding space just like words. States of the parallel GRU layers are later concatenated and passed into another GRU layer with attention mechanism (context layer). The output of this layer is then passed into a softmax layer, which outputs the probability of the punctuation mark to be placed between the current and the previous word (starting from the second word in sequence).

The attention mechanism is useful for the neural network to identify positions in a sequence where important information is concentrated (Bahdanau, 2014). For words, it helps to focus on positions of words and word combinations that signal the introduction of a punctuation mark. For prosodic features, it either remembers a salient point in the sequence or detects a certain movement that could help determining a punctuation mark at a certain position.

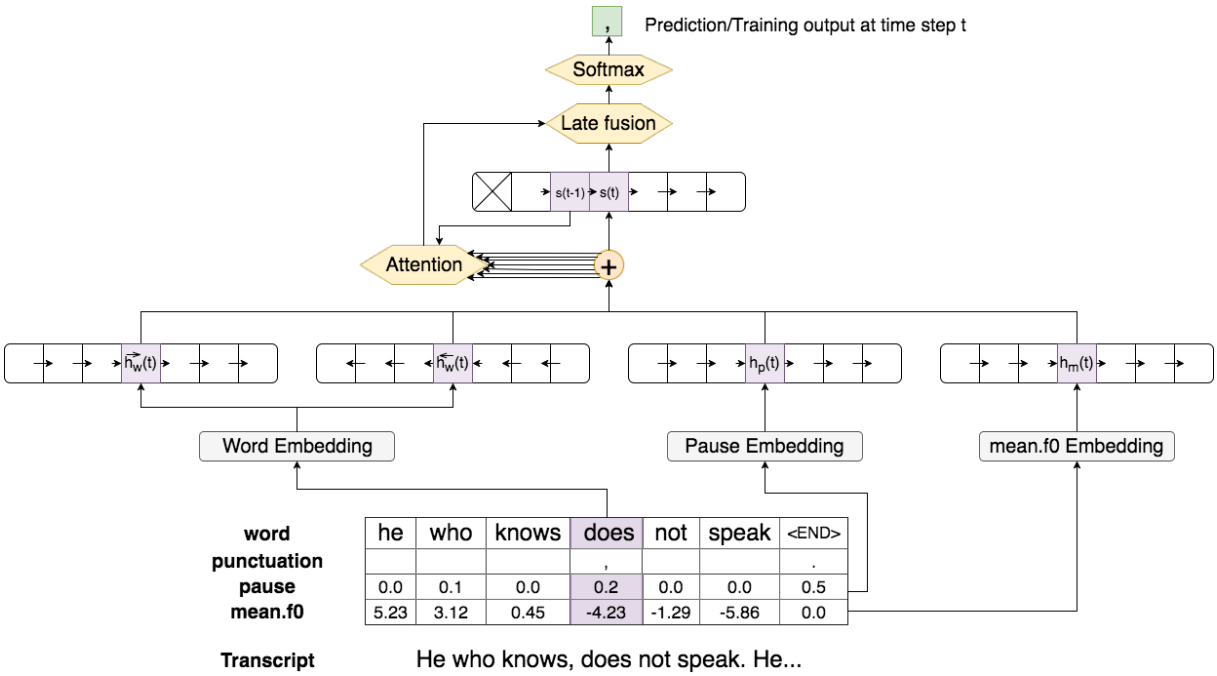

Figure 5. Attention neural network architecture depicting the processing of a speech data sample with pause and mean f0 features aligned at the word level.

The model was implemented using a Theano library and was first tested on a prosodically annotated corpus containing TED talks (Farrús et al., 2016). The corpus consists of 1046 talks by 884 English speakers, uttering a total of 156034 sentences. Acoustic-prosodic features are calculated on word and sentence levels using word-alignment information. F0 and intensity measurements are converted to semitones relative to speaker means value for normalization.

Taking into account that the number of words per sentence in our corpus is 15-20 in average, the data is sampled into sequences of size 50. 59811 samples were extracted this

way. 70% percent (41867 samples) of this data was allocated for training, 15% for testing and 15% for validation (8971 samples each).

Three main prosodic features were tested: pause durations between words, fundamental frequencies of words (mean and range) and intensities of words (mean and range). Range values were calculated as the difference of the maximum and minimum for each word.

The best results in our tests were obtained when pause and f0 mean features were used. This setting gave an F-score of 65.7% in correct punctuation placement in the testing set. This is an improvement of 9.1% over the baseline setting by Tilk et al. For detecting periods, the best feature set includes pauses, mean f0 and range intensity (76.2% F-score). More details on experimental setting and elaborated results are presented in Öktem (2017).

3.4.2 Integration into the KRISTINA pipeline

The experimental setup presented previously is separated into two modules: (1) prosodic feature extraction, and (2) punctuation generation. The first module has been built into a module called *Proscripter*, and takes the voice input and its aligned transcription and produces an acoustically enriched transcription. This representation of speech called *Proscript* is then fed into the punctuator module and punctuated text is obtained.

This setup should be placed right after the text-to-speech processing, so the subsequent modules will benefit from the punctuated text. Both the voice input and word alignments (in the right format) are needed for the punctuation module to work.

3.5 Language-specific system development

The speech-to-text technologies existing at the beginning of the project target two types of data: broadcast speech or conversational telephone speech. Neither is a good match with the KRISTINA data. In order to improve recognition accuracy, the acoustic and language models have been adapted to the KRISTINA domain. Word recognition accuracy is a key issue to avoid misleading the classification decisions taken by the language analysis modules. The following sections summarize the language specific adaptation work performed in KRISTINA.

3.5.1 Language modeling

Texts from different sources were used to build language models for Spanish, German, Polish, Turkish and Arabic. The amount of training data available is summarized in Table 9.

Source	Spanish	German	Polish	Turkish	Arabic
KRISTINA Web	24M	780k	-	8.2M	-
KRISTINA dialogues and transcriptions	3.6k	50k	4.2k	1.2k	12.7k
Other (transcriptions and Web)	2.7G	2.2G	574M	591M	1.5G

Table 9. Number of words after text normalization available for language model training. Units: k=x10^3, M=x10^6, G= x10^9.

A portion of KRISTINA audio transcriptions and simulated dialogues was included into the language model training corpus. Textual data from the Web sources inventoried by KRISTINA partners were downloaded, cleaned and normalized. Websites like *doctissimo*, which aims to "*help everyone proactively manage their health and well-being by providing free, clear and*

reliable information along with forums to share (your) experiences", were found very useful for adapting the speech-to-text systems. Several themes are treated in this website like health, psychology, pregnancy, nutrition, sexuality, among others. The questions found in the forums seem to be close to what a KRISTINA user would ask, as for example, in Spanish:

- *¿Es seguro tomar ácido fólico?*
- *¿Es peligroso hacer ejercicio durante el embarazo?*
- *¿cuantos kilos habeis engordado durante el embarazo?*

Other sites are reliable from the point of view of content, but the language is distant from conversational speech. They helped in reducing the OOV in all languages, but were less important in improving perplexity of the test data. A separate component language model was estimated only on the questions automatically extracted from the available data for all languages. This component LM was expected to fit well the KRISTINA data in terms of speech style and use cases.

Since some sites contain information in several languages, a character-based language identification algorithm was used. Information retrieval techniques were evaluated (Zhang et al., 2015). The goal was to rank documents based on a fitting criterion and select only the most relevant for language model training. The gains obtained with such a technique were limited.

Morphological decomposition (Virpioja et al., 2013) was applied to the morphologically rich languages, German, Turkish and Polish, in order to increase language model coverage. The following example, accompanied by an English translation, illustrates the morphological structure of the Turkish language:

ev = *house*
evim = *my house*
evin = *your house*
evde = *in the house*
evdekiler =*people in the house*
evdekilere = *to people in the house*

Decomposition led to improvement in terms of speech recognition accuracy in German.

Other broadcast and conversational speech transcriptions, as well as other Web texts available prior to KRISTINA were used in language modeling. Such non-related KRISTINA data constitute around 99% of the training data on each language.

Several component language models were built for each language. They were interpolated to fit the likelihood of the language specific development data sets, comprised of audio transcriptions for KRISTINA recordings. The development data is disjoint from the training corpus, and contains about 2 hours of speech for each language. Although small in terms of volume (less than 1%), the KRISTINA data account for between 20% and 40% of among the other LM components. This highlights the importance of having reliable data matching the target domain for language modeling.

Table 10 shows the out-of-vocabulary rate and the perplexity obtained with the language models from the baseline system and with the language model tailored for KRISTINA in four languages. The comparison in Arabic has been omitted from the table since the models have

not been released yet. The baseline models are those available at the beginning of KRISTINA and target broadcast data. LM adaptation improves the OOV rate of development data in all languages evaluated. The highest improvement was obtained on German, for which perplexity decreased by 51% relative. Generally, perplexity gains of more than 10% relative can be considered significant (Rosenfeld, 2000).

Language	Baseline		KRISTINA	
	OOV(%)	ppx	OOV(%)	ppx
Spanish	0.39	217	0.22	164
German	1.43	499	0.53	246
Polish	1.71	840	1.29	664
Turkish	4.91	830	2.8	538

Table 10. Out-of-vocabulary rate (OOV) and perplexity (ppx) of the KRISTINA development data calculated using the baseline language model and the one adapted to KRISTINA.

Acoustic and pronunciation modeling

The acoustic models were built using state-of-the-art discriminative training methods. Models for Spanish, Polish and Arabic are estimated using broadcast acoustic data available prior to KRISTINA. Several hundreds of hours of annotated data are used in the training process. Turkish acoustic models are estimated on both broadcast speech and conversational telephone speech data, together with their manual transcriptions. Besides the original acoustic data, Turkish models make use of synthetically created samples, as described in Section 3.3

The German acoustic models were trained on hundreds of hours of manually transcribed data and more than a thousand hours of untranscribed data. The untranscribed audio consists of podcasts gathered from a variety of radio and TV shows from Germany, Austria and Switzerland. Automatic transcriptions were generated for the untranscribed, with which unsupervised acoustic model training was done. A method similar to Lileikyte et al. (2016) was applied.

The phone sets cover the language-specific phones and special units to model silence, breath and filler words. Acoustic units are linked to words via pronunciation dictionaries. In this work, they are obtained using linguistically derived grapheme-to-phoneme (g2p) rules.

The pronunciation lexica are obtained to cover various accents, which is one of the main challenges into today's speech recognition technologies (Benzeghiba et. al, 2007). The influence of accent variation was conducted in a multi-accented Spanish corpus. Spanish is written using a Latin alphabet and has 27 letters. The most particular characters are the *ñ* and accentuated vowels (*á, é, í, ó, ú*). Spanish orthography is overall regular, i.e. there is almost one-to-one correspondence between letters and phonemes. This one-to-one mapping is kept in the pronunciation dictionary in addition with some extra rules to treat spelling difficulties.

Several experiments were conducted to improve the system performance. For each version of pronunciation dictionary, an acoustic model was trained and the accuracy measured on several multi-accented development data sets. The pronunciation dictionary was modified to

take into account pronunciation varieties from different regions in Europe and South America. The accuracy varies by about 2.5% relative depending on the test data. The short/long vowel differentiation makes word error rate varying on less than 0.8% relative. Alternative pronunciations representing coda /s/ and intervocalic /d/ lenition were incorporated into the lexicon, obtaining a 1.8% variation into the word error rate.

Post-processing

The punctuation module currently integrated in KRISTINA is based on statistical language models. Colon, period, and question marks are automatically added to the transcription hypothesis by means of a punctuation model. Syntactical-based rules were added to the punctuation system to detect question utterances. The rules are based on language-specific question words, for instance in Spanish, *cuándo*, *donde*, *quién*, etc. Such techniques were applied to Spanish, German, Polish and Turkish.

The post-processing system also capitalizes utterances. Numbers are kept in their orthographic form. Post-processing parameters were tuned on the KRISTINA development data, with special attention given to stop marks.

3.5.2 Summary of ASR system evaluation

During system development, ASR systems are iteratively evaluated. Table 11 summarizes the WER results obtained on the KRISTINA data with the main system releases. Intermediate systems are omitted in the comparison, as well as results in Arabic. The baselines are the broadcast recognition systems available at the start of KRISTINA, which operate on batch mode. The second set of models was released at M18 for the first prototype, in which Spanish, German and Polish systems are real-time. The M18 systems have state-of-the-art acoustic models and lightly adapted language models. The third and fourth sets of models were released at M26 and M32 respectively. They were obtained after improving acoustic model training methods and adapting the language models further to the KRISTINA domain. The M32 Turkish acoustic models were obtained using acoustic data augmentation methos, as described in Section 3.3.

Language	Baseline	M18	M26	M32
Spanish	28.4*	22.9	22.1	22.1
German	44.2*	38.3	35.5	30.3
Polish	38.0*	33.7	26.4	20.6
Turkish	57.1*	48.9*	38.9	33.7

Table 11. WER (%) of the different ASR system releases on KRISTINA development data. Values marked with a star (*) represent batch systems.

The adaptation of speech recognizers to the KRISTINA domain led to relative gains between 22% and 41% relative. The word error rates of the M32 systems range between 20% and 34% on the KRISTINA data. For comparison, WERs range between 10% and 20% on broadcast speech conversations.

The largest gains were obtained on Turkish, which is still the most challenging language in the KRISTINA project. The system outputs were analyzed by a native speaker and compared to

the ground truth transcriptions. In most of cases, recognition errors are due to differences in word suffixes, for example:

Example 1:

Reference:	babamla tanıştığında ve hamile **kaldığında** işi bıraktı.
Translation:	*She stopped working when she met my father and **when** she got pregnant.*
Hypothesis:	babamla tanıştığında *ben* hamile **kaldığından** işi bıraktı.
Translation:	*She stopped working when she met my father and **because** she got pregnant.*

Another source of discrepancy between the reference and recognition hypothesis is the (incorrect) use of grammar by non-native speakers. Word order in Turkish sentences is generally subject-object-verbe, but speakers living in foreign countries (Germany in this case) may use different orders, resulting in recognition errors. Here is an example:

Example 2:

Reference:	Beraber oturuyoruz dışarıda onunla.
Hypothesis:	Beraber oturuyoruz.
Standard grammar:	Onunla beraber dışarıda oturuyoruz.
Translation:	*We sit outside together with her.*

4 SPOKEN LANGUAGE UNDERSTANDING

In this section, we report on the advances made for Task 3.3 that addresses the analysis and capture of the natural language user utterances into structured, ontological representations, so that appropriate system responses can subsequently be inferred. This includes multilingual corpus annotation (see Section 2.4), the development of multilingual dependency parsers and of lexical resources, and the definition of a framework for the projection of the extracted dependency-based linguistic representations into ontological ones. We also report on the adaptation to spoken language and on the communicative dimension of the content. The analysis pipeline comprises the following steps:

- Tokenization, Part-of-speech tagging, lemmatization;
- Surface-syntactic parsing (SSynt structures);
- Mapping to deep-syntactic structures (DSynt structures);
- Mapping to language-independent abstract structures (Conceptual structures).

The KRISTINA text analysis framework adheres to a multi-layer paradigm: starting from the hypothesized transcribed user utterances, representations of higher abstraction are successively obtained, until the underlying semantics are distilled in a formal and language-independent manner that allows for automated reasoning and interpretation. Figure 6 shows the representations corresponding to the respective layers, namely surface-syntax, deep syntax (in the original language and in English), conceptual and ontological, for the example sentence “¿Cuál es la temperatura ideal del agua para bañar a un bebé?”, lit. “What is the ideal temperature of the water to bathe a baby”, which is encountered in the baby care domain of the Use Case 2, Health Expert scenario.

.

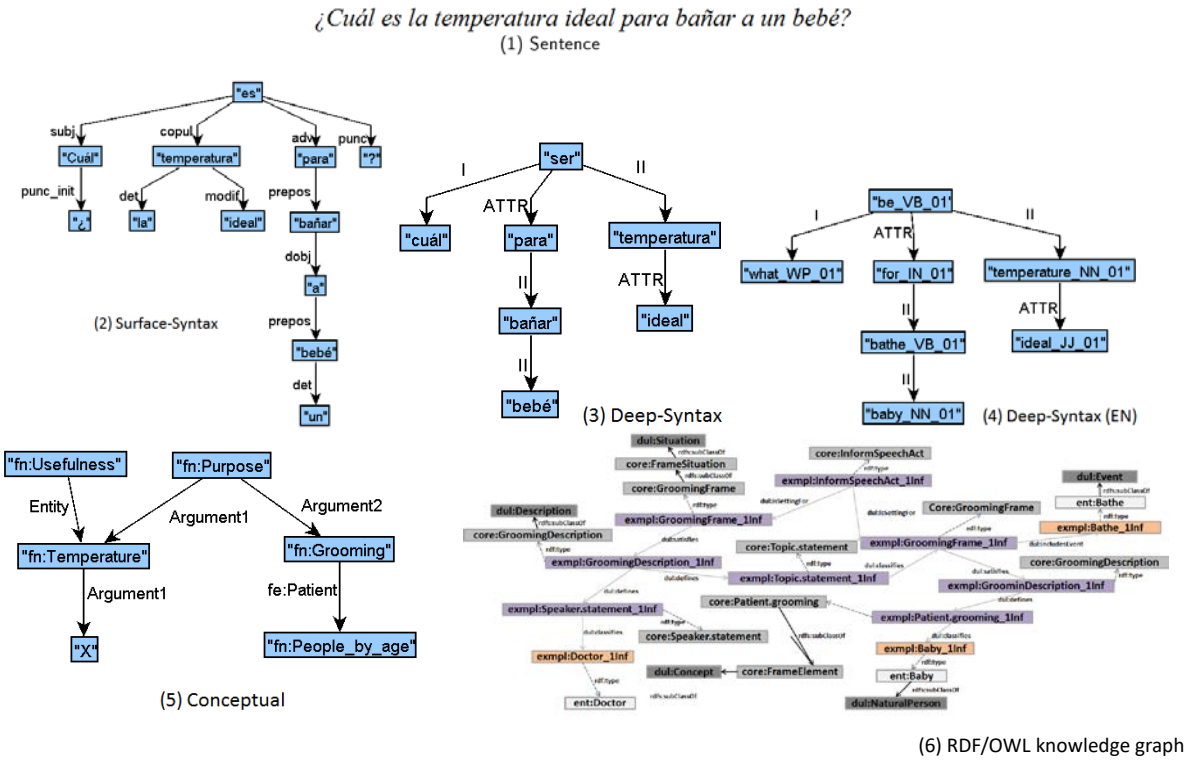

Figure 6. Successive representations of the example sentence "¿Cuál es la temperatura ideal del agua para bañar a un bebé?" obtained by the KRISTINA multi-layer text analysis.

4.1 Lexical resources

Lexical resources are crucial for appropriately handling a multilingual analysis pipeline. The dictionaries for KRISTINA have been constructed having in mind one main goal, namely create a resource that can be used for both analysis and generation purposes, keeping in mind the following constraints: (i) the lexicons are part of a multilingual pipeline, that is, they need to be interconnected in order link the concepts in the different languages and thus enable linguistic abstraction in a language-independent way; (ii) the lexicons are constrained by the grammatical resources (graph-transduction grammars) that use the information contained in the lexicons during the analysis process. Behind differences between languages, all of them contain information for both generic entries (Figure 7) and specific –lexical-entries.

```
nounExtArg_:_predicateExtArg_ {
        pos = "NN"
        spos = "noun"
        gp = {
              I = {
                dpos = N
                rel = obl_compl
            }
```

```
                    II = {
                       dpos = N
                       rel = obl_compl
                     }
                  }
               }
```

Figure 7. Generic entry for Spanish nouns with an external argument, according to Nombank.

As we can observe in Figure 7, each generic entry reveals relevant equivalences that are inherited by specific entries (except if some information is overwritten in a specific entry). Such equivalences can embed other equivalences, and so on. For instance, in Figure 7, the information for each actant (I and II) -expressed through equivalences regarding morphological and syntactic information- are embedded within the first-order equivalence for "gp" (the government pattern).

The mapping by-default between the corresponding SSynt and the DSynt relations, as well as the conversion between the PoS automatically generated by the parser and a unified PoS tagset, are also included in the dictionary, although each mapping is enriched with the corresponding morphological information. As the lexicons are used for analysis and generation at the same time, they also include a small section of mapping between argument-predicate relations (A0, A1, etc.) and DSynt relations.

Regarding the specific lexical entries (see Figure 8 as example), each lexicon is composed by all the lexical entries included in the KRISTINA dialogues.

```
"lesen_VB_01":_verb_ {
      entryID = "3"
            vnID = "learn.14-1"

      Eng_tr = "read"
      pbID = "read.01"
      fn = "Reading"
      onID = "read-v.1"
      lemma = "lesen"
      gp = {
            I = {
                  pos = "NN"
                  case = "nom"
                  ssyntrel = "SB"
            }
            II = {
                  pos = "NN"
                  case = "nom"
                  ssyntrel = "OA"
            }
               }
   }
```

Figure 8. Entry for the German verb lesen 'to read'.

Each entry has a unique corresponding entry ID and, depending on its PoS, different information is included. For nouns, we include the following attributes: lemma, countable (with values "yes" or "no"), Nombank ID (if available), Framenet ID (if available), OntoNotes ID (if available), English_translation, and morphological information about its governed elements (PoS, presence of a specific adposition, case and/or finiteness). Depending the language, gender is also included. [3] For verbs, the attributes included are: lemma, Propbank ID (if available), Verbnet ID (if available), Framenet ID (if available), OntoNotes ID (if available), English_translation, and morphological information about its governed elements (PoS, case, presence of a specific adposition, SSynt relation used). For other PoS, just the lemma and the English_translation attributes are included across languages, although relevant information is added depending on the language (e.g. in German, prepositions entries include the case assigned by the preposition).

In order to manage ambiguity, each sense of a word is entered as an independent entry. That is why a number is contained in the "name" of each entry (01 in the example).

4.2 Syntactic and semantic parsing

4.2.1 Surface-syntactic parsing

For surface-syntactic parsing, a new dependency parser (Dyer et al. 2015; Ballesteros et al. 2016) and control structure for sequence to sequence neural networks that allows to model stack like structures has been developed. In addition, character-based representations of words have been explored; the idea behind was to use recurrent neural network to capture morphosyntactic clues replacing standard look-up based word representations by orthographical representation of words. This implied statistical sharing across word forms that are similar on the surface (Ballesteros et al. 2016) and improvement in morphologically rich languages. The parser we developed is very fast (since it runs in a single core), light (it only requires 1GB of RAM memory) and more accurate when the same resources are used. The code of the parser can be found at https://github.com/clab/lstm-parser/tree/char-based. A sample dependency parse in Polish is shown in Figure 9.

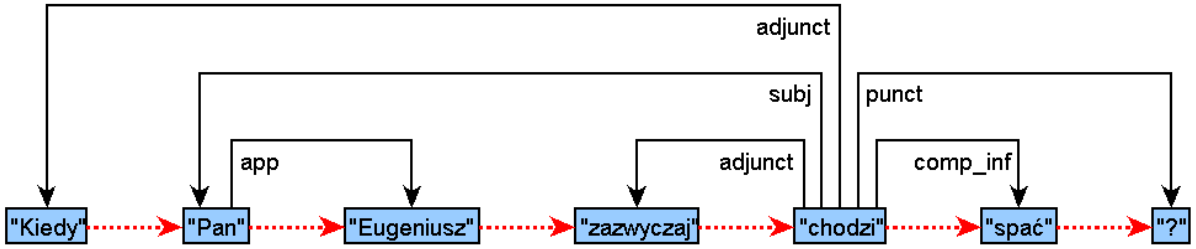

Figure 9. *A sample Polish SSyntS: "When does Mr Eugene usually go to bed?".*

We adapted the new parser to spoken language: character-based representations that are able to handle out-of-vocabulary words (Ballesteros et al. 2016) have been applied to the spoken language corpus and led to an improvement of about 4 points over the baseline.

[3] If the gender of the entry is not intrinsic to the entry but depends on morphology (e.g. Spanish "gato/gata" for 'cat'), the masculine is assumed as default entry.

4.2.2 Deep-syntactic parsing

In order to abstract from language-specific features of the aforementioned dependency parsers, we also aim at structures in which only content-bearing words are present, and semantics-oriented relations between them are made explicit. For this, first-in-their-genre multilingual rule-based and statistical deep-syntactic transducers have been developed.

The objective of this kind of transducer is to identify and remove all functional words (auxiliaries, determiners, void adpositions and conjunctions), and to generalize the syntactic dependencies obtained during the previous stage, while adding sub-categorization information for syntactic predicates. We have two options for transducing surface-syntactic dependency trees into deep-syntactic trees: a statistical system and a rule-based system. The statistical transducer (Ballesteros et al. 2014), trained on parallel SSynt and DSynt corpora (see, for instance, Mille et al. 2013 for Spanish), has been developed for English and Spanish. The rule-based transducer consists of graph-transduction grammars that access to language-specific lexicons to remove the void prepositions and conjunctions, when any is available (Mille & Wanner 2015), and assign predicate-argument edge labels between the remaining words. We have rule-based transducers for English, Spanish, German, Polish and Turkish. In Figure 10, the deep-syntactic structure corresponding to Figure 9 is shown; edge labels are now oriented towards semantics instead of syntax. In such a simple Polish sentence, only the question mark is removed, there are no functional elements. The code for the statistical deep-syntactic transducers and rule-based transducers is provided in the Github repository (cf. footnote 4 and 5, respectively[4] [5]).

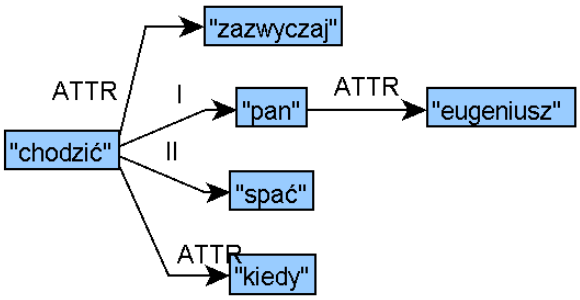

Figure 10: A sample Polish DSyntS corresponding to Figure 9*.*

4.2.3 Towards semantic parsing: the problem of multilinguality

Since an ontological representation is completely independent from linguistic considerations in general, it is even more so language-independent. Hence, the language analysis pipeline must be able to handle knowledge expressed in different languages in the same way. For mapping deep-syntactic structures to more abstract, language-independent linguistic representations, large scale lexical resources are needed. Unfortunately, such resources are only available in English at this point, for example PropBank (Kingsbury & Palmer 2002), FrameNet (Baker et al. 1998), VerbNet (Schuler 2005) and the mappings between them (SemLink, Palmer 2009). For this reason, we chose to map all input languages to English. For

[4] https://github.com/talnsoftware/deepsyntacticparsing

[5] https://github.com/talnsoftware/DSynt_Converter

all languages, after the SSynt-DSynt transduction, all idiosyncratic words have been left out, and only meaningful ones were kept in the structure. In other words, the parallelism between the deep-syntactic representations of different languages is such that substituting word labels of a language X to English word labels produces a (very often) accurate English deep-syntactic structure. Using multilingual resources such as BabelNet, it will possible to obtain the translations of these words into English on a large scale. For the first prototype, we manually crafted and linked lexical resources for the lexical units that are used in the dialogues of the different use cases.

Figure 11 shows a node label substitution between Polish and English. The structure is perfectly isomorphic to the Polish DSyntS, and is a valid English DSyntS. Note that the original Polish sentence, which contains all but only the words of the DSyntS, is quite different to its English counterpart, which would contain more nodes: *"When **does** Mr Eugene go **to** sleep?"*.

Deep-syntactic structures are then normalized and simplified, in order to make further processing easier; see for instance the DSyntS in Figure 12. The mappings described in this section are performed with manually crafted graph-transduction grammars.

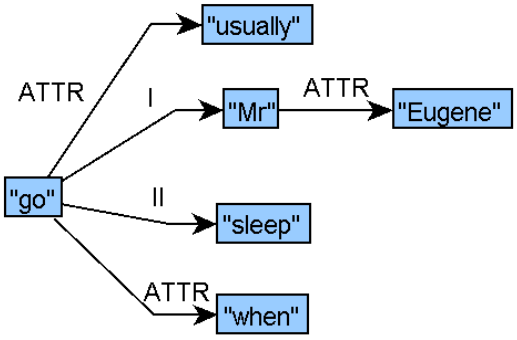

Figure 11: A sample English DSyntS obtained from the Polish DSyntS in Figure 10.

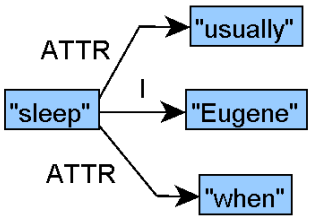

Figure 12: A sample normalized DSyntS corresponding to Figure 11.

4.2.4 Mapping to ontological representations

Having abstracted away from language-specific idiosyncrasies, the next step is to transform the predicate-argument-based representations of the input utterances, as afforded by deep syntactic parsing, into formal structured representations upon which reasoning can take place in order to understand the conversation context and deduce appropriate system responses (see WP5).

As explained in Deliverable D5.2, where, among others, the state of the art in the formalization of lexical semantics is discussed along with the requirements pertinent to the KRISTINA conceptual infrastructure, our approach adheres to the semantic frame-based paradigm: the input utterances' semantics are captured by means of n-ary relational contexts that describe events/situations (aka, semantic frames), the involved elements (aka, frame participants) and their respective semantic roles (aka, frame roles). In line with Semantic Web ontology design practices and standards, and the seminal work by (Gangemi 2010), which delineates frame semantics in view of the Descriptions and Situations (DnS) ontology pattern, we opted for a reified representation, interpreting frames as *dul:Descriptons*, frame roles as *dul:Concepts* and the extracted relational contexts as *dul:Situations*. For example, a sentence such as *"Mr Eugene goes for a walk at noon."* is interpreted as a situation that involves the activity *"going for a walk"*, where the participating elements *"Eugene"* and *"noon"* are classified as (i.e. serve the roles of) the agent and the temporal attribute respectively of the referred activity context.

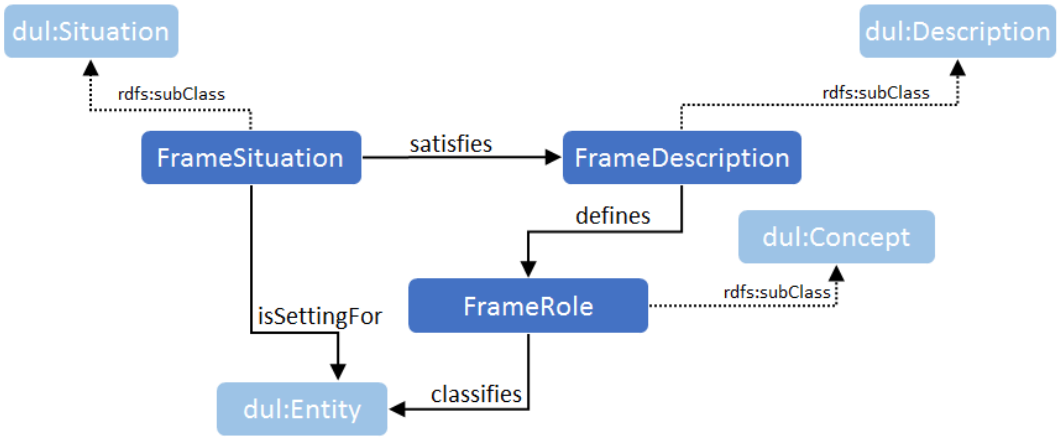

Figure 13. Frame situation ontology.

Figure 13 shows the core reference model classes:

- **FrameSituation.** It captures *relational contexts* on a set of entities, i.e. events and situations describing *n-ary* relations between entities (e.g. the liking of having a glass of milk before going to bed). We distinguish between EventSituations, i.e. relational contexts pertinent to events/activities (e.g. going to bed, watching TV) and AttributeSituations, i.e. those pertinent to attributive descriptions (e.g. favorite board game, feeling sad).
- **FrameDescription.** It encapsulates the frame-based descriptive contexts (i.e. conceptualisation) that define the interrelations (semantic roles) of the set of participating entities.
- **FrameRole.** It serves as a superclass for the concepts used to classify the entities specified by frame descriptions, by means of their semantic roles. The subclasses (e.g. Agent, Theme, etc.).

Abstracting away from the specifics and premises underlying the various predicate-argument linguistic resources (namely PropBank, VerbNet, FrameNet (Giuglea & Moschitti 2004)), the

specializations of the FrameRole class draw upon a generic set of semantic roles that mediate the different semantic role definitions of the various linguistic resources, while catering for the scope and expressivity needs within the KRISTINA use case domains, thus resulting in the following definitions:

- Context: captures the referenced event/activity or attribute
- Agent, Theme, Destination: capture agentive, patientive and locative information respectively
- TemporalAttribute: captures temporal information pertinent to a situation, including start and end time, as well as duration
- FrequencyAttribute: captures information about the occurrences of a situation with respect to a given period
- TemporalPattern: captures temporal relations between situations, including ordered and overlapping intervals.

Along the same lines, the pertinent frames (i.e. the descriptive contexts of the recognized events and situations) are not grounded to those of a specific reference resource, allowing greater flexibility when it comes to mapping to the reference domain ontologies. As far as negation is concerned, we opted for a representation with RDF-flavoured semantics, using a data type property to annotate the referenced relational context with information about whether its value is true or false.

In addition to the utterance contents, the resulting ontological representation captures also its performative function (speech act). This is accomplished by drawing upon the KRISTINA dialogue act ontology, which specifies the different types of dialogue acts (e.g. statement, request, acknowledge, etc.) and provides properties for associating dialogue acts instances with their respective contents.

To accomplish the transition from the deep syntactic structures to ontological representations, a semantic structure representation layer is used as pivot. Our semantic structure follows the principles of the Meaning-Text Model. This allows us to preserve the grounding to the initial linguistic structures[6], while affording a principled projection to respective OWL statements. The defined semantic structure representation enriches the node-edge typology of the deep syntactic layer by introducing additional node and edge types that allow the typing of the predicate-argument structures in alignment with the afore-described model. Examples include *role* nodes (e.g. "Time", "Duration", "Frequency", "Range"), *temprel* nodes (e.g. "Temporal_Ordering", "Temporal_Overlap"), and *speech act* nodes (e.g. Request), and pertinent edges (e.g. "part/whole", "topic").

Continuing the running example of previous subsection (i.e. "*When does Mr Eugene usually go to bed?*"), Figure 14 shows the resulting semantic structure (further annotated with VerbNet labels when available).

[6] Thereby, providing also support for the inverse transformation, i.e. the generation of NL expressions starting from ontological representations.

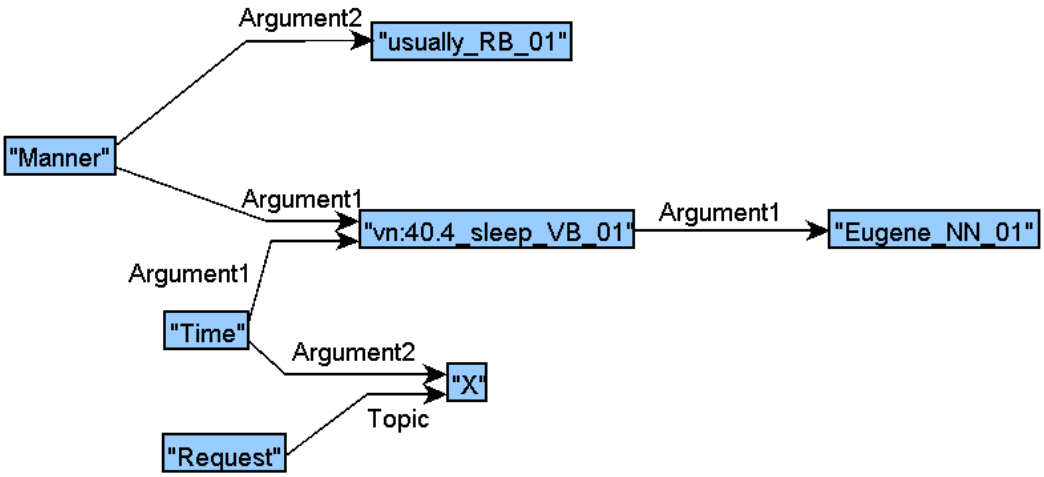

Figure 14. The sample semantic structure corresponding to *Figure 13*.

An excerpt of the corresponding OWL representation (represented in Turtle notation) is shown in the following:

```
:sleepCtx1 a dul:Situation .
  context:includesEvent :sleep1 ;
  context:includes :Eugene , time1, :usually1 ;
  context:satisfies :sleepDesc1 , :requestDesc1 .
:sleepDesc1 a context:Description ;
  context:defines :c0, :c1, c2, c3 .
:c0 a context:Context ;
  context:classifies :sleep1 .
:sleep1 a onto:Sleep .
:c1 a context:Agent ;
  context:classifies :Eugene .
:c2 a context:TemporalAttribute ;
  context:classifies :time1 .
:time1 a onto:Time .
```

4.3 Communicative analysis techniques

Communicative analysis in the framework of KRISTINA is grounded in the theoretical framework of Information Structure whose origin goes back to Mathesius (1929), and which is also known as Topic-Focus Articulation (TFA) (Sgall, 1967) in the Prague School (Daneš, 1970) and Communicative Structure in the Meaning-Text Theory (Mel'čuk, 2001), determines the "communicative" segmentation of the meaning of an utterance. The MTT establishes different levels of abstraction, and information or communicative structure pivots the transition between semantic and syntactic levels. Communicative structure reflects speakers' particular arrangement of the content of a sentence, based upon what is known content and what is new information about that content. Mel'čuk (2001) distinguishes several dimensions of communicative structure (CommStr), among which thematicity specifically addresses the distinction between the theme (what the statement is about) and the rheme (what new information is provided), and includes a third span, the specifier, that sets the context of the sentence. A further key concept in Mel'čuk's theoretical framework is hierarchical thematicity described over propositions rather than sentences. This implies that thematicity spans can

be embedded within spans and one sentence can contain several propositions at the same level or embedded.

In the context of the communicative analysis technologies sub-task, the annotation guidelines for thematicity are still under development. Previous work in this area was limited to written language (Bohnet et al., 2013b), consequently, extension has been necessary given that spoken language analysis is required within the KRISTINA project. In spoken language, spans may occur that are pure dysfunctionalities (which are typical of spoken language), which somehow must be recognized and filtered from the communicative assignation. At the same time, some well-established criteria that can be applied to grammatical sentences need to be adapted to sentences that are not always grammatical.

On the other side, several theoretical studies argue that morphology, syntax, and information (or communicative) structure, that organizes a given content with respect to the intention of the speaker, show a strong correlation with intonation. It has been argued in the literature that: (i) prosody expresses the communicative intention of the speaker (Grice, 1989); (ii) the communicative intention of the speaker is to a large extent encoded in terms of the Information Structure (Steedman, 2013); (iii) Information Structure is rendered both through syntax and prosody (Mel'čuk, 2001); (iv) in Content to Speech (CTS) applications, the Information Structure of a sentence can be derived in a content organization procedure, as done in Natural Language Text Generation (NLTG) (Wanner et al., 2003; Bouayad-Agha et al., 2012). However, little empirical work based on sufficiently large corpora has been carried out so far to buttress this argumentation.

To provide further empirical insights on this issue, an analysis on the correlation between communicative structure and prosody has been carried out on spoken language (voice samples used in these experiments were described in Section 2.4) in order to study if the previously reported conclusions on read speech (Domínguez et al., 2014) can be applied to spoken language as well. In our previous empirical work (Domínguez et al., 2016a), we report for read speech a characteristic rising ToBI pattern in the theme (either L*+H pitch accent or L-H% boundary tone or a combination of them), a falling ToBI pattern in the rheme (H* L-L%) and a rising pith accent (L*+H) or a flat contour (L* L-L%) in the specifier span.

For the annotation of prosodic phrases and ToBI labels, we have deployed the implementation of an automatic prosody tagger (Domínguez et al., 2016c), which carried out a segmentation based on acoustic features from raw audio following our previous empirical work in this direction (Domínguez et al., 2016b) including a combination of fundamental frequency (F0), intensity and duration parameters to mark boundaries at the prosodic phrase level.

The analysis carried out on spontaneous monologues confirms the tendency outlined above for theme and specifier spans. Table 12 summarizes the results on the thematicity spans coinciding with prosodic phrases, which are 50% of the total amount of 328 prosodic phrases. Results include all three thematicity spans, i.e. theme (T1), rheme (R1) and specifier (SP1) as well as backgroundedness (B) (another dimension of communicative structure according to Mel'čuk), and proposition spans, where we distinguish between single propositions, same level (or coordinated) propositions and embedded propositions.

As can be observed in the highlighted figures of Table 12, T1 coinciding spans are associated with a rising tune in 88% of the examples, and SP1 spans also present a tendency of having

either a rising (64%) or a flat (36%) pattern. However, R1 matching spans to prosodic phrases show a majority of 44% in rising patterns. This can be explained due to the fact that in spontaneous monologues, speakers tend to use a continuation rise (L-H% boundary tone) to indicate that they will continue talking. This analysis also helps to broaden the picture provided in previous analysis of read speech in two directions: spoken language contains more examples of backgroundedness than read speech and propositions are a communicative unit, which is found to be matching to prosodic phrases.

ToBI Pattern	Themacity				Proposition		
	T1	R1	SP1	B	Single	Same Level	Embedded
Rising	88%	44%	64%	31%	43%	48%	42%
Falling	13%	38%	-	31%	30%	34%	36%
Flat	-	21%	36%	38%	26%	18%	21%

Table 12: Results on the correlation between communicative structure and prosody.

4.4 Evaluation

4.4.1 Surface- and deep-syntactic analysis

Table 13 presents surface-syntactic parsing results obtained using MaltParser (Nivre, 2007) with its default settings (Bohnet et al., 2013)[7] and (Ballesteros et al.,2016), which has been developed within KRISTINA. MaltParser, our baseline, is one of the first statistical dependency parsers ever developed. It produced state-of-the-art results until 2009 and, since then, has been used as a reference parser in the litterature (see for example the shared task on parsing morphologically rich languages (Seddah et al., 2013)). Both (Bohnet et al., 2013) and (Ballesteros et al., 2016) produce state-of-the-art results. (Ballesteros et al., 2016) apply recent developments in recurrent neural networks, which makes it faster and better for parsing out-of-vocabulary words.

We present results in the reference benchmarks from CoNLL 2009 treebanks (Hajič et al. 2009) for the languages used in the project in terms of labeled attachment score (LAS). The English treebank does not have morphological features. The results for (Bohnet et al., 2013) were obtained using morphological features for the remaining languages. Such features were explicitly used only for Polish, Turkish and Arabic with the Ballesteros et al. (2016)'s parser.

Parser	English	German	Spanish	Polish	Turkish	Arabic
MaltParser, default settings	86.0*	80.7	82.4	75.6	65.7	80.4
(Bohnet et al., 2013) (integrated)	90.6*	89.4	89.6	81.8	70.9	82.9
(Ballesteros et al., 2016)	89.9*	88.2*	88.0*	82.7	70.9	83.8

*Table 13: Surface dependency parsing results: labeled attachment score (%). Results marked with * imply that they were obtained without explicit morphological features.*

[7] https://code.google.com/p/mate-tools/wiki/ParserAndModels

Both, (Bohnet et al., 2013) and the KRISTINA parser (Ballesteros et al., 2016), outperform the MaltParser baseline for all languages, and they approach or even surpass the best results ever published in the litterature for these languages. For English, the improvement over the baseline (MaltParser) is of more than 3 points, what corresponds to a 37.5% error reduction in terms of LAS. For German, the improvement is even higher, especially with the Bohnet et al. (2013)'s parser (+8 points, -49% error rate); KRISTINA parser (+7, -43% error rate) achieves a very competitive performance even without the use of explicit morphological features. For Spanish, the trend is similar: improvements of 6 points (-42% error rate) for (Bohnet et al., 2013) and 5 points (-34% error rate) for the KRSITINA parser.

In terms of speed performance, KRISTINA parser is 4 times faster than Bohnet et al. (2013)'s parser. The KRISTINA parser uses a greedy decoding strategy. The parser is trained to minimize cross-entropy relative to a distribution of gold-standard sequences, obtained by transforming labeled syntactic trees via a manually defined procedure. At testing time, the parser makes greedy decisions according to the learned model. Bohnet et al. (2013)'s parser uses a beam search approach, which explores larger searching spaces, and, consequently, is slower (4 times) and consumes more memory (1GB for the KRISTINA parser *vs.* 32GB for (Bohnet et al., 2013)).

The results provided by our parsers (often state-of-the-art, see the KRISTINA parser, show they are ready to be used on a large scale project like KRISTINA. To demonstrate that this is the case, we assessed our parsers on the manually annotated Spanish spoken data. Table 14 shows the results obtained by training our parser on the KRISTINA Spanish data. From a total of 700 sentences, a 100 were selected for evaluation (there are 1228 dependencies in the test set). The KRISTINA parser outperformed Bohnet et al. (2013)'s parser by more than 4 points in terms of LAS.

Parser	LAS	UAS	LA
Bohnet et al. (2013)	69.6	83.8	72.8
Ballesteros et al. (2016)	73.9	84.0	78.2

Table 14: Results of the adaptation of the Spanish parser to spoken language (%). LAS: labeled attachment score; UAS: unlabeled attachment score; LA: label accuracy.[8]

For deep-syntactic (DSynt) parsing, we carried out evaluations of the systems we developed so far on English, Spanish and German. There is no baseline so far for deep-syntactic parsing. In order to evaluate the DSynt parsing in German and French, 51 sentences have been manually annotated with deep-syntax in each language (942 and 664 words respectively), and compared to the annotation produced by our rule-based transducers. For Spanish, a gold standard evaluation set of 258 sentences (5,641 words) was available; we also provide evaluations for English (1,299 semi-supervised annotated sentences for evaluation, 42,480

[8] 'LAS', i.e., "labeled attachment score", is the ratio between the number of tokens of the obtained syntactic tree for which their "head" (or governor) function and the corresponding dependency relation are correctly recognized and the total number of head nodes in the tree. 'UAS', i.e., "unlabeled attachment score", is the ratio between the number of correctly identified head – dependent pairs and the total number of the head – dependent pairs in the tree. 'LA', i.e., "label accuracy", is the ratio between the number of correctly identified dependency relation labels and the total number of dependency relations in the tree.

words). For these last two languages, statistical transducers have been developed in addition to rule-based ones. Two aspects were evaluated:

Hypernode identification evaluation:

$$F1h = \frac{2\text{ph} \cdot rh}{ph + rh}$$

where

- ph is the number of correctly predicted nodes divided by the total number of predicted hypernodes
- rh is the number of correctly predicted hypernodes divided by the number of hypernodes in the gold standard.

Dependency labels evaluation:

- Unlabeled attachment precision (UAP): number of nodes with a correctly predicted governor divided by the total number of predicted nodes.
- Labeled attachment precision (LAP): number of nodes with a correctly predicted governor and governing relation label divided by the total number of predicted nodes.
- Unlabeled attachment recall (UAR): number of nodes with a correctly predicted governor divided by the total number of gold nodes.
- Labeled attachment recall (LAR): number of nodes with a correctly predicted governor and governing relation label divided by the total number of gold node.

Table 15 shows the evaluation results for the SSynt-DSynt transitions only in order to not take into account errors produced by the surface-syntactic parsers.

Measure	Spanish (ML)	English (ML)	Spanish (RB)	English (RB)	German (RB)
F1h	99.51	98.88	97.31	98.12	97.71
LAP	91.07	90.63	79.57	86.97	89.01
UAP	98.32	93.70	88.95	90.77	92.72
LAR	90.57	91.02	83.25	89.08	86.60
UAR	97.78	94.11	93.07	92.97	90.21

Table 15: Deep-dependency parsing results labeling (%). ML: Machine Learning based parser; RB: Rule based parser.

Feature	Language	Prototype 1	Prototype 2
Languages supported		DE, ES, PL	DE, ES, PL, TR
Number of rules (SSynt-DSynt)	DE	148	379

	ES	226	216
	PL	177	237
	TR	-	653
Number of rules (DSynt-Con)	ALL	488	783
Number of sentences supported	DE	24	389
	ES	25	246
	PL	32	130
	TR	-	117
Number of lexical units in lexicon	DE	-	385
	ES	21	324
	PL	113	244
	TR	-	

Table 16. Comparison between P1 and P2 analysis modules.

At this point, Prototypes 1 and 2 have been delivered, and Prototype 3 is being developed. Table 16 summarizes the work underpinning the extended domain coverage, elaborating the extensions to the lexicons and the rules that cater for domain coverage and the mappings across the pairs of representation layers respectively. As the expressivity of the frame-based modelling and mapping of the conceptual representations to respective ontological ones already covered the considered semantics (i.e. events/situations and participating entities, temporal relations, frequency and negation), there has been no need for further refinements.

Table 17 shows the domains covered within the scenarios addressed in the first and the second prototypes, outlining the extended coverage provided by text analysis.

Language	Prototype 1	Prototype 2
DE	Social companion: weather; newspaper; Nursing assistant: sleeping habits	Social companion: weather; newspaper; social media news; local events Nursing assistant: sleeping habits; eating habits; sleeping problems, diseases Health expert: dementia; diabetes; recipes; sleep hygiene

ES	Health expert: - Baby care: child's room, bathing, sleeping; - Back pain: sciatica	Health expert: - Baby care: child's room, bathing, sleeping, child's growth and breastfeeding, vaccination and side effects, recommended activities; - Back pain: sciatica, symptoms, causes, treatment Mediator: - Baby care: recommended child's checkups, vaccination schedule; - Back pain back pain: suggestion according to the BMI, exercises for low back pain tension relief Receptionist: information about NGOs, health care centres, appointments with GO and specialists
PL	Nursing assistant: sleeping habits	Nursing assistant: sleeping habits; eating habits; sleeping problems, diseases Health expert: dementia; diabetes; recipes; sleep hygiene
TR		Social companion: weather; newspaper; social media news; local events

Table 17. Domains covered by text analysis in the first and second prototypes.

4.4.2 Mapping to ontological representations

The quality of the transformation of the semantic predicate-argument structures to respective ontological representations is manifested in terms of its ability to capture the semantics underlying the user utterances so that subsequently appropriately responses can be deduced from the knowledge base. An indirect way to evaluate how accurately the semantic representations of the user utterances capture the underlying meaning is by examining the obtained system responses; though undoubtedly useful for assessing the overall performance, such evaluation depends also on the performance of the query matching approach, thus blurring insights into how the individual tasks perform. Aiming at a more informative assessment, we focus instead directly on the expressiveness afforded by the resulting representations with respect to the information types that need to be accommodated. Towards this end, we compare the expressiveness of the resulting knowledge graphs with that of two state-of-the-art approaches, namely FRED (Gangemi et al., 2016) and PIKES (Corcoglioniti et al., 2016) and discuss the implications of the different modeling choices, focusing on the requirements as delineated by the use cases addressed within KRISTINA, namely the capturing of speech act types, temporal attributes and relations, frequency information, and negation.

All three approaches follow a frame-based representation where events and situations are represented as reified objects connected to their participants by means of explicating their semantic roles. FRED considers primarily frames expressed by verbs and uses VerbNet and

FrameNet for frame and semantic role labeling; in addition to VerbNet and FrameNet, PIKES uses also ProbpBank for labeling frames pertinent to verbs, and considers also frames expressed by nouns, using NomBank for labelling. Moreover, both employ coreference resolution and role propagation for merging equivalent individuals and entity linking for enriching the graphs with links to existing Semantic Web resource, including DBpedia. The approach implemented in KRISTINA considers frames expressed by verbs or nouns and follows an ontology pattern-based design, similar uses in its current stage a VerbNet-like labeling scheme for roles, while performing a In the following we discuss how the three approaches under comparison, namely FRED, PIKES and the KRISTINA one, accommodate the different.

Speech acts

Developed within a dialogue system context, the KRISTINA approach models explicitly the speech act type of the input NL sentence, currently distinguishing between requests (i.e. wh-/how- and yes/no questions) and statements. The speech act type is captured through a distinct individual of the SpeechAct class that serves as a container for aggregating the situations comprising the contents of the input text. In addition, a named individual is introduced as a placeholder for the requested element, currently typed as a member of either the DomainEntity class (the top-level level class of the domain ontology) or the Event class, in cases where it can be deduced without ambiguity that the referred entity is an event. Thus, given the sentence “When does Eugene go to bed?” we would have the following triples:

```
:spechAct rdf:type :Request .
:sleep context:Agent :Eugene ;
        context:TemporalAttribute :entity .
:entity rdf:type onto:Time .
```

Neither FRED nor PIKES cater for the speech act type of the analysed sentence. Yet, in case of what-questions, FRED models the interrogative pronoun as an instance of owl:Thing, while where- and when-questions trigger the introduction of Location or Unit_of_Time individuals respectively. PIKES on the other hand, tends to maintain the question auxiliary (e.g. “do” in “what does he like to eat?”) and exhibits higher volatility in the recognition of the interrogative pronouns semantic typing.

Temporal attributes

These include the time at which an event or situation with a temporal extension happens and the duration of the referenced event or situation. In the KRISTINA implementation, a TemporalAttribute class is used for semantic role labelling, while the class type (i.e. Time or Duration) of the labelled entity serves for further refinement; the actual values (e.g. “midnight” , “6 hours”) are asserted through a datatype property. Consequently, the sentences “Eugene sleeps at midnight.” and “Eugene sleeps six hours.” would result in among others the following triples:

```
:sleep context:Agent  :Eugene ;
        context:TemporalAttribute :time1 .
:time1 rdf:type onto:Time ;
        onto:hasValue “midnight”.
```

```
:sleep context:Agent  :Eugene ;
        context:TemporalAttribute :dur1 .
:dur1 rdf:type onto:Duration ;
        onto:hasValue “six hours”.
```

FRED and PIKES follow a less transparent and explicit representation, relying primarily on subsequent processing for the interpretation of the intended semantics. Temporal prepositions are in their majority captured in FRED as respectively labelled properties in its custom namespace; for instance, the preposition “at” in “Eugene usually sleeps at midnight.” results in a property assertion of the form: fred:sleep fred:at fred:midnight. Temporal adverbs (e.g. today) on the other hand, are linked to the referenced situation using the dul:associatedWith property (which is used in several other cases with very different meaning), thus also without making explicit the underlying temporal semantics. Leaving out the occasional labelling through dedicated PropBank argument labels (e.g. as in the case of the frame “last.01”), PIKES resorts in most cases to the PropBank “AM-TMP” role for typing the participation link between the temporal entity and the referenced situation, using FrameNet’s Time frame element for further specialisation in the cases of temporal adverbs.

Temporal relations

In the current implementation we consider two types of temporal relations, namely temporal order that encodes the “before”/ “after relations in Allen’s interval algebra and temporal overlap that encodes respectively the “contains”/”during” relations. As in the case of temporal attributes, the representation of the semantic role is performed in a combined manner through a TemporalPattern class, used to capture the nature of the semantic role, and the class of the labelled entity that explicates the type of the pattern. Similar to the case of temporal attributes, the representations of FRED and PIKES follow a less transparent paradigm. In FRED, custom properties (e.g. fred:before, fred:during) are again introduced to link the pertinent elements, while the focus on primarily verbally expressed events and situations limits the possibilities of exploiting, when available, more specialised labels afforded by FrameNet. Although PIKES does not share this limitation and FrameNet-based typing for temporal relations would be expected, it uses instead the ProbBank “AM-TMP” role, thus falling short to effectively distinguish between the different temporal relations, as well as between temporal attributes and temporal relations.

Negation

FRED and KRISTINA follow the same approach for the representation of negation, namely an RDF-oriented form, where the referenced situation is annotated with the information that its truth value is false. PIKES on the other hand, applying graph refinements mainly for materialisation and elimination of redundant entities rather than towards a more ontology-design compliant representation, uses the PropBank “AM-NEG” role label to link the referenced frame to the element expressing negation (e.g. not, never, etc.).

Frequency

As activities of daily living and pertinent routines form an important aspect of the domains addressed by the KRISITNA use cases, capturing frequency information is essential. In the

current implementation, a special class is introduced for representing frequency patterns and encoding the rate and referenced period time. For example, the query “How often does Eugene go to the toilet at night?” and its response “He goes to the toilet between 1 and 4 times.” would result, among others, in the following set of triples respectively:

```
:go context:Agent  :Eugene ;
        context:FrequencyAttribute :freq1 ;
        context:Destination :toilet1 ;
    context:TemporalAttribute :night1 .
:freq1 rdf:type onto:Frequency .
```

```
:go context:Agent  :Eugene ;
        context:FrequencyAttribute :freq1 ;
:freq1 rdf:type onto:Frequency ;
        onto:hasValue :val1 .
:val1 rdf:type onto:Range ;
        onto:hasStartValue “1” ;
    onto:hasEndValue “4” ;
```

FRED uses the dul:hasQuality to capture the relation between an entity and its qualities. Thus for instance, given the sentence “Eugene owns a black cat.”, we would have the triple <fred:cat dul:hasQuality fred:black>. Adverbs denoting frequency (e.g. always, often, etc.) are treated as qualities of the referenced event (situation), and as such captured likewise through the dul:hasQuality property. Implicit and more complex frequency expressions through (e.g. “three times a night”) on the other hand, are simply treated as any other type of information. PIKES, supporting adverbial frames too, affords in general the semantic typing of adverbs of frequency; for complex frequency expressions through, it resorts to the the PropBank “AM-TMP” role, which being also used to capture time, duration and temporal relation information, aggravates ambiguity.

5 WORK TOWARDS THE FINAL DEMONSTRATOR

This section summarizes the ongoing work planned to be accomplished by the end of the lifetime of the project.

During the remaining period of the project, the Arabic system developed to KRISTINA will be released for evaluation in the final demonstrator. Acoustic data augmentation will be applied, in priority, to Arabic and German to increase speech recognition robustness. The neural network based punctuation generation module will be integrated into the KRISTINA pipeline. The models will be trained and tested for English as a control language, as well as for the Spanish KRISTINA data. An exhaustive evaluation of accuracy and speed will follow in order to integrate the punctuation system.

We plan at expanding the dictionaries for covering the dialogues of the final system. We will also focus on ensuring uniformity across languages, and on optimizing the access to the lexica by the graph-transduction rules, so as to improve processing speed.

We will ensure the coverage of the sentences for the final prototype by the graph-transduction grammars. If necessary, we will train new deep-syntactic parsers on the multilayered data annotated in the framework of the project (e.g in Polish).

Currently, the automatic derivation of thematicity from syntax is being investigated. Initial experiments in German with retrieved text from the sleeping hygiene scenario are being carried out to develop a rule-based approach for the derivation of thematicity. The goal of these experiments is to implement a reliable tool for the analysis of web retrieved texts, upon which the generation of expressive prosody (described in WP6) based on the analyzed thematicity-prosody correspondence can be built.

The experiments in German are based on the syntactic parsing in CONLL format using the Bohnet's parsers trained on the TIGER corpus. The referent unit is the word. The following criteria are being taken into consideration for the detection of propositions and thematicity spans:

- Position in the sentence
- Part of speech
- Dependency function
- Dependency relation with respect to other words in the sentence

The derivation of level 1 and level 2 thematicity is foreseen, based upon the application of these spans for the derivation of prosody as reported in (Domínguez, 2017). Both rule-based and machine learning approaches are being investigated. In the final version of the system, there will be a module for the analysis of hierarchical thematicity from web retrieved text. This module will parse the text on the first place, and derive hierarchical thematicity thereupon. Hierarchical thematicity will be used in the final system for the derivation of more communicative and, thus, expressive prosody by the language generation module.

6 CONCLUSIONS

This deliverable summarized the advances on the vocal analysis technologies in the context of KRISTINA. It focused on the development of multilingual automatic speech recognition (ASR) and multilingual spoken language understanding (SLU) technologies, which aim, respectively, to transform the user speech signal into text, and text into ontological structured representations for processing by the subsequent modules of the KRISTINA platform.

Developing and evaluating such technologies require a large amount of annotated data matching the real usage conditions. For evaluation purposes, the audio data collected by the user partners was manually or semi-automatically annotated in terms of word transcriptions, speaker turns and morphological, surface-syntactic and deep-syntactic structures. Such data was used to drive development of statistical models and to evaluate the systems.

The ASR systems have been adapted in order to improve the recognition accuracy on the KRISTINA data, and to reduce the latency delay as required for a real-time dialogue applications. Specific types of acoustic models and specific decoding strategies were designed to ensure a real-time system response: for instance, the Spanish recognizer achieved a latency of about 0.44 seconds in average. Real-time acoustic models have been developed for the five KRISTINA languages. Acoustic data augmentation was assessed as a means to reduce the mismatch between training and testing conditions. Data augmentation contributed to reducing the word error rate on Turkish from 38.9% to 33.7%.

Language-specific development was carried out to adapt the systems to the KRISTINA domain. For each of the languages, improved acoustic modeling techniques have been deployed, and tailored language models have been integrated. As a result, compared to the baseline systems available at the start of KRISTINA, relative word error rate improvements ranging from 22% (in Spanish) to 46% (in Polish) were obtained. The word error rates obtained attest that the KRISTINA task is particularly challenging. For comparison, depending on the language, word error rate performance is generally around 8-15% for well-trained broadcast recognition systems. The error rates can easily double for tasks targeting spontaneous conversational speech or in degraded acoustic conditions.

The core of multilingual dependency parsers was developed during the first half of the project, but they have been adjusted, complemented and expanded during the second half. The surface-syntactic analyzer developed within KRISTINA was shown to be highly competitive. Compared to the baseline: it is more efficient in terms of time and memory; it obtained similar labeled attachment scores (LAS) on the six languages assessed; and a 4% absolute gain in terms of LAS on the KRISTINA Spanish data. Statistical and rule-based deep-syntactic parsers were developed. A framework for the projection of the extracted dependency-based linguistic representations into ontological ones is also described, showing the progress and the improvements within the KRISTINA project.

7 REFERENCES

Atalay, Nart B., Kemal Oflazer, and Bilge Say. (2003). The annotation process in the Turkish treebank. In *Proc. of the 4th Intern. Workshop on Linguistically Interpreted Corpora (LINC)*.

Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In CoRR.

Baker, CF., Ch. J. Fillmore and J.B. Lowe. (1998). The Berkeley FrameNet project. In *Proc. of the 17th international conference on Computational linguistics,* Volume 1, Association for Computational Linguistics, 86–90.

Ballesteros, M., B. Bohnet, S. Mille and L. Wanner (2014). Deep-Syntactic Parsing. In *Proceedings of COLING (COLING 2014)*. Dublin, Ireland.

Ballesteros, M., C.Dyer, Y. Goldberg and N.Smith. (2016). Greedy Transition-based Dependency Parsing with Stack-LSTMs . *Computational Linguistics*. MIT Press.

Benzeghiba, M. et al (2007). Automatic speech recognition and speech variability: a review, Speech Communication, Volume 49, Issues 10-11.

Bohnet, B., J. Nivre, I. Boguslavsky, R.Farkas, F. Ginter and J.Hajic. (2013). Joint Morphological and Syntactic Analysis for Richly Inflected Languages, *Transactions of the Association for Computational Linguistics*.

Bohnet, B., A. Burga, and L. Wanner. (2013b). Towards the Annotation of Penn TreeBank with Information Structure, in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan. pp. 1250–1256.

Bouayad-Agha, N., G. Casamayor, S. Mille, and L. Wanner. (2012). Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges, *ACM Transactions on Speech and Language Processing*, vol. 9, no. 2.

Brants, S., S. Dipper, P. Eisenberg, S. Hansen-Schirra, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on language and computation* 2, no. 4: 597-620.

Daneš, F. (1970). One instance of Prague School methodology: Functional analysis of utterance and text, *Garvin*, pp. 132–141.

Deng, L., Acero, A., Plumpe, M., and Huang, X. (2000). Large-vocabulary speech recognition under adverse acoustic environments. In *Proc. ISCA Interspeech*.

Domínguez, M., M. Farrús, A. Burga, and L. Wanner. (2014). The Information Structure-Prosody Language Interface Revisited, in *Proceedings of the 7th International Conference on Speech Prosody* (SP2014), Dublin, Ireland, pp. 539–543.

Domínguez, M., M. Farrús, A. Burga, and L. Wanner. (2016a). Using hierarchical information structure for prosody prediction in content-to-speech applications. In *Proceedings of the 8th International Conference on Speech Prosody*, (SP2016), Boston, USA, pp. 1019–1023.

Domínguez, M., M. Farrús, and L. Wanner. (2016b). Combining acoustic and linguistic features in phrase-oriented prosody prediction. In *Proceedings of the 8th International Conference on Speech Prosody*, (SP2016), Boston, USA, pp. 796–800.

Domínguez, M., M. Farrús, and L. Wanner. (2016c). An Automatic Prosody Tagger for Spontaneous Speech. *Under submission*

Farrús, M, Lai, C & Moore, J 2016, Paragraph-based Prosodic Cues for Speech Synthesis Applications. in Proceedings of Speech Prosody 2016.

Fousek, P., Lamel, L., and Gauvain, J.-L. (2008a). On the use of MLP features for broadcast news transcription. *Text, Speech and Dialogue, Lecture Notes in Computer Science*, 5246:303–310.

Fousek, P., Lamel, L., and Gauvain, J.-L. (2008b). Transcribing broadcast data using MLP features. In *Proc. ISCA Interspeech*, pp. 1433–1436.

Gales, M. (2001). Multiple-cluster adaptive training schemes. In *Proc. IEEE ICASSP*, volume 1, pp. 361–364.

Garrido, J.M., Escudero, D., Aguilar, L., Cardeñoso V., Rodero E., de-la-Mota C., González C., Vivaracho C., Rustullet S., Larrea O., Laplaza Y., Vizcaíno F., Estebas E., Cabrera M., Bonafonte A. (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. In Lang Resources & Evaluation 47: 945.

Gangemi, A. (2010). What's in a Schema? *C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, and L. Prevot, editors, Ontology and the Lexicon*. Cambridge University Press.

Gauvain, J.L., Gangolf, J.J., and Lamel, L. (1996). Speech recognition for an information kiosk. In *Proc. International Conference on Speech and Language Processing*, pp. 849-852.

Gauvain, J. L. (2001). Audio Partitioning and Transcription for Broadcast Data Indexation. *MTAP Journal* , 187-200.

Gauvain, J.-L., Lamel, L., and Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, 37:89–108. 31, 43

Grézl, F., Karafiát, M., Kontár, S., and Cernocky, J. (2007). Probabilistic and bottle-neck features for LVCSR of meetings. In *Proc. IEEE ICASSP*, volume 4, pp. 757–761.

Grice, P. H. (1989). Further Notes on Logic and Conversation, in *Studies in the Way of Words*, pp. 41–57.

Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M.A. Martí, L. Màrquez, A. Meyers et al. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1-18. Association for Computational Linguistics.

Hermansky, H., Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, vol. 2, num. 4, pp. 578-589.

Hermansky, H. (1990). Perceptual linear prediction (PLP) analysis for speech. *Journal of the Acoustical Society of America*, 87:1738–1752.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proc. IEEE ICASSP*, volume 1, pp. 181–184.

Johansson, R., and Nugues, P. (2007). Extended constituent-to-dependency conversion for English, In *Proceedings of 16th Nordic Conference of Computational Linguistics*, p. 105-112.

Kingsbury, P., and Palmer, M. (2002). From Treebank to PropBank, In *Proceedings of the Third International Conference on Language Resources and Evaluation*, p. 1989-1993.

Liao, H., and Gales, M. J. F. (2005). Joint uncertainty decoding for noise robust speech recognition. In *Proc. ISCA Interspeech*, pp. 3129-3132.

Lileikyte, R., Gorin, A., Lamel, L., Gauvain, J.-L., Fraga-Silva, T. (2016). Lithuanian broadcast speech transcription using semi-supervised acoustic model training, In *Proc. SLTU 2016*.

Lippmann, R., Martin, E. and Paul, D. (1987). Multi-style training for robust isolated-word speech recognition. In *Proc. IEEE ICASSP*, vol. 12.

Mangu, L., Brill, E., and Stolcke, A. (1999). Finding consensus among words: lattice-based word error minimization. In *Proc. European Conference on Speech Communication Technology*, pp. 495–498.

Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mathesius, V. (1929). Zur Satzperspektive im modernen Englisch, In *Archiv für das Studium der neueren Sprachen und Literaturen*, 155, pp. 202–210.

Mel'čuk, I. (1988). *Dependency syntax: Theory and practice*, State University of New York Press.

Mel'čuk, I. (2001). *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*.

Mille, S., A. Burga, and L. Wanner. (2013). AnCora-UPF: A Multi-Level Annotation of Spanish. In *Proceedings of the 2nd International Conference on Dependency Linguistics (DepLing)*.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G.Eryiğit, S.Kübler, S.Marinov and Marsi, E. MaltParser. (2007). A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95-135.

Öktem, A., Farrús, M., Wanner, L. Attentional Parallel RNNs for Generating Punctuation in Transcribed Speech. In 5th International Conference on Statistical Language and Speech Processing, Le Mans, France.

Palmer, M. (2009). SemLink: Linking Propbank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference (GenLex-09)*, Pisa, Italy.

Ragni, A., Knill, K., Rath, S. and Gales, M. (2014). Data augmentation for low resource languages, In *Proc. ISCA Interspeech*, pp. 810-814.

Rosset, S., Bennacef, S., and Lamel, L. (1999). Design strategies for spoken language dialog systems. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, pp. 1535-1538.

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? In *Proceedings of the IEEE*, 88(8):1270–1278.

Seddah D et al. (2013). Overview of the SPMRL 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Shared Task on Paring Morphologically Rich Languages*, EMNLP 2013, Seattle, U.S.

Schuler, K.K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*, Ph.D. thesis, University of Pennsylvania.

Schwarz, P., Matjka, P., and Cernocky, J. (2004). Towards lower error rates in phoneme recognition. In *Proc. International Conference on Text, Speech and Dialogue*, pp 465–472.

Sgall, P. (1967). Functional Sentence Perspective in a generative description of language, in *Prague Studies in Mathematical Linguistics*, vol. 2, pp. 203–225.

Steedman, M. (2013). The surface-compositional semantics of English intonation, In *Language*, vol. 90, pp. 2–57.

Taulé, M., M.A. Martí, and M. Recasens. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC*.

Tilk, O. and Alumäe, T. (2016). Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In Proceedings of Interspeech.

Veselý, K., Ghoshal, A., Burget, L., and Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *Proc. ISCA Interspeech,* pp. 2345-2349.

Virpioja, S., Smit, P., Grönroos, S.A., and Kurimo, M. (2013). Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline, *Aalto University publication series*, ISBN: 978-952-60-5501-5.

Wanner, L., B. Bohnet, and M. Giereth. (2003). Deriving the Communicative Structure in Applied NLG, In *Proceedings of the 9th European Workshop on Natural Language Generation at the Biannual Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 100–104.

Wolinski, M., K. Glowinska, and M. Swidzinskil. (2011). A Preliminary Version of Skladnica - a Treebank of Polish. In *Proceedings of the 5th Language & Technology Conference, Poznan*, pp. 299-303.

Cui, X., Goel, V., and Kingsbury, B. (2015). Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(9), 1469-1477.

Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., and Tsakalidi, S. (2015). Enhancing Low Resource Keyword Spotting with Automatically Retrieved Web Documents, In *Proc. ISCA Interspeech*, pp. 839-843.