# The Development of the Multilingual LUNA Corpus for Spoken Language System Porting

**Evgeny A. Stepanov, Giuseppe Riccardi, Ali Orkan Bayer**

Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Italy
{stepanov,riccardi,bayer}@disi.unitn.it

**Abstract**

The development of annotated corpora is a critical process in the development of speech applications for multiple target languages. While the technology to develop a monolingual speech application has reached satisfactory results (in terms of performance and effort), porting an existing application from a *source* language to a *target* language is still a very expensive task. In this paper we address the problem of creating multilingual aligned corpora and its evaluation in the context of a spoken language understanding (SLU) porting task. We discuss the challenges of the manual creation of multilingual corpora, as well as present the algorithms for the creation of multilingual SLU via Statistical Machine Translation (SMT).

**Keywords:** Multilingual Corpora, Spoken Language Corpora, Parallel Corpora

## 1. Introduction

Speech services are becoming increasingly spread (e.g. call centers, smart-phones, etc.). The common limitation of the most available speech services is the lack of multilingual support: the services are developed only for the languages with rich available resources (usually English). Consequently, the large user bases of speakers of other languages are left out. The main reason for this is the fact that developing the same speech service in another language is an expensive manual effort; since it requires additional data collection and annotation. An alternative is an automatic cross-language porting of an existing service to another language via translation. However, it has severe data resource limitations. (1) Available multilingual resources such as aligned corpora are few in number and are different from conversational data in style. (2) Annotation in most monolingual language resources, such as Penn Treebank, is designed for linguistic analysis and hardly suitable for building data-driven spoken language systems. (3) Few existing parallel spoken conversation corpora represent resource-rich or close family language pairs.

There are very few parallel spoken conversation corpora specifically designed for building data-driven spoken language systems. The available ones are either translated to close languages (e.g. PORTMEDIA: French - Italian (Lefèvre et al., 2012)), or to English (e.g. ATIS: English - Chinese (He et al., 2013)). Multilingual LUNA Corpus, on the other hand, is the translation of Italian LUNA Corpus via professional translation services that covers both close (Spanish), and distant family languages (Turkish and Greek). [1] Thus, it allows for broader perspective on cross-language system portability. At the same time, it allows to address issues of cross-language porting differences to linguistic resource-rich and resource-poor languages.

We first describe the source data – Italian LUNA corpus, and then specifics of the translation of conversation transcriptions. Then, we present the use case – cross-language SLU porting task Multilingual LUNA corpus was used in.

[1] The corpus is available for research purposes upon signing Data Sharing Agreement with University of Trento.

## 2. Italian LUNA Corpus

The Italian LUNA Corpus (Dinarelli et al., 2009) is a collection of 723 human-machine dialogs (approximately 4,000 turns & 5 hours of speech) in the hardware/software help desk domain. The dialogs are conversations of the users involved in problem solving collected using Wizard of Oz (WOZ) technique: the human agent (wizard) reacting to user requests is following one of the ten scenarios identified as most common by the help desk service provider. Text-to-Speech Synthesis (TTS) was used to provide responses to the users. The dialogs are organized in transcriptions and annotations defined within FP6 LUNA Project.

### 2.1. Levels of Annotation

The dialogs were annotated at different levels: words, turns, attribute-value pairs, predicate argument structure and dialog acts:

- The annotation at word level consists of lemmas, part-of-speech tags and morpho-syntactic information following EAGLES corpora annotation (Leech and Wilson, 1996).
- Attribute-value annotation makes use of predefined ontology of domain concepts and their relations.
- Predicate argument annotation is based on FrameNet model (Baker et al., 1998).
- Dialog act annotation was inspired by DAMSL (Core and Allen, 1997), TRAINS (Traum, 1996), and DIT++ (Bunt, 2005) and is used to mark intentions in an utterance.

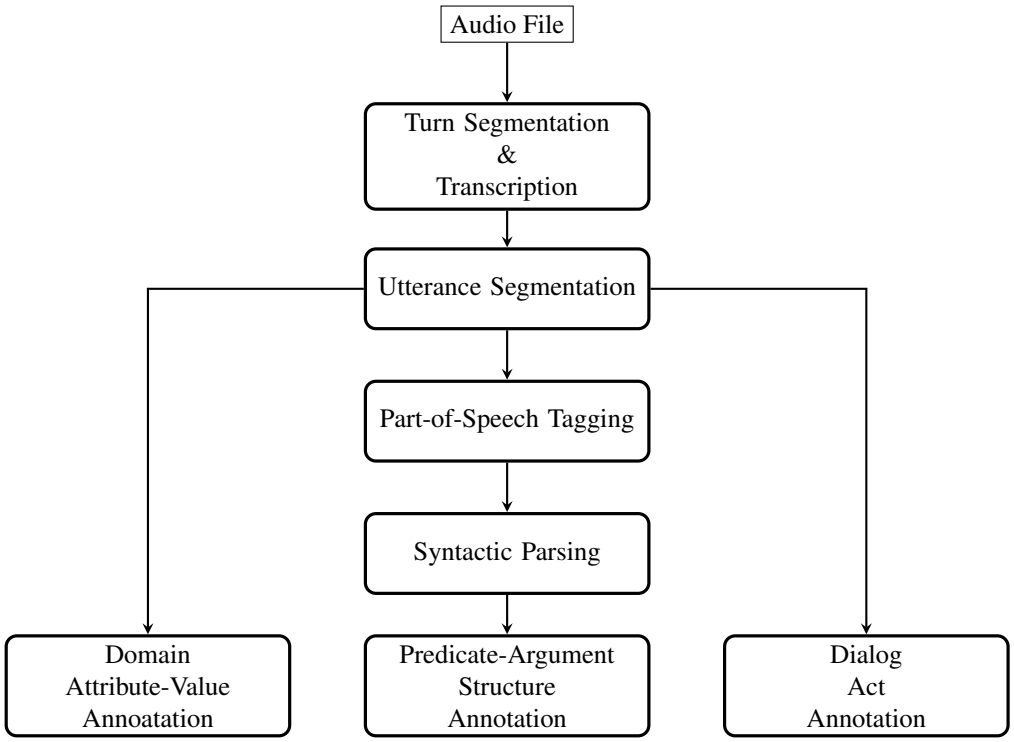


Figure 1: Italian LUNA Corpus annotation process (from (Dinarelli et al., 2009)).

| Annotation Level | # of dialogs |
|---|---|
| Attribute-Value | 723 |
| Dialog Act | 224 |
| Predicate-Argument Structure (FrameNet) | 129 |

Table 1: Statistics on LUNA Corpus annotation levels

The general process of annotation can be seen on Figure 1. Dialog act and attribute-value annotation is done on segmented dialogs at utterance level. However, predicate argument annotation requires POS-tagging and syntactic parsing. This was done semi-automatically using the Bikel parser trained on an Italian corpus (Corazza et al., 2007) with subsequent manual correction. Different levels of annotation cover different subsets of the corpus. Table 1 provides information on the amount of dialogs annotated at each level.

### 2.2. Anonymization

Within FP7 PortDial Project LUNA Human-Machine Corpus has gone through additional process of anonymization. Sensitive private information, such as personal names, phone numbers were replaced with random values: named entities were replaced with a random named entity of the same type drawn from a list of common Italian entities. and phone numbers were replaced with a random numeric sequences. A special attention was given to preserve the distribution of token frequencies within anonymized concept values.

Additionally, an automatic Spoken Language Understanding (SLU) model was trained and tested on anonymized data to ensure that the step has no significant impact on the performance. We used a popular approach for Spoken Language Understanding models – Conditional Random Fields (CRFs) (Lafferty et al., 2001) – which model the conditional probability of the concept sequence given the word sequence.

The LUNA Spoken Language Understanding model – the baseline model in (Bayer and Riccardi, 2013) – uses the following features:

- *Orthographic*: first and last $n$ letters of a token, where $n$ ranges from 1 to 5 (10 features);
- *Ngrams*: unigrams and bigrams of tokens in the window of $\pm 1$ tokens, including $-1, 1$ token pair (6 features);
- *Binary*: a feature to label numerical expressions (1 feature);

All the features are independent in the window of $\pm 1$ tokens. Additionally, CRFs use previous output token as a feature for current token decision.

The SLU model trained on original LUNA Corpus has Concept Error Rate (CER) of 21.5%, and the model trained and tested on anonymization corpus has CER of 21.7% (the difference is insignificant).

## 3. Manual Creation of Multilingual Corpora: Professional Translation

Within the FP7 PortDial project, the Italian LUNA Human-Machine Corpus (all 723 dialogs) has been translated by expert translators to Spanish, Turkish and Greek. The translated corpus consists of text only (i.e. annotations have not been transferred); and is intended as a reference resource

| IT | ES | TR | EN |
|---|---|---|---|
| *ciao Paola* | *hola, Paola,* | *merhaba Paola* | *hi Paola* |
| *ho un problema* | *tengo un problema* | | *I have a problem* |
| ***con la sta*** *con la tastiera* | ***con la pe...*** *con el teclado,* | *klavye ile* ***[disf="klavye ile"]*** | ***with the pri[nter]...*** *with the keyboard* |
| | | *bir sornum var,* | |
| *ho un tasto staccato mi è rimasto in mano e* | *tiene una tecla pegada, se me ha quedado en la mano y* | *bir tuş çıktı, elimde kaldı ve* | *I have a button off that remained in my hand and* |
| ***non*** *non riesco più a usarla* | ***no*** *no consigo usarla* | *artık kullanamıyorum* ***[disf="kullanamıyorum"]*** | ***cannot*** *use it anymore* |

Table 2: An example of speech disfluency translations in a single utterance from Italian (IT) to Spanish (ES) and Turkish (TR). English translation (EN) is given for reference only.

for research on data-driven spoken language system porting. In this section we describe the process and challenges associated with *manual* creation of multilingual conversational corpora.

### 3.1. Transcribed Speech Translation Artifacts

Since the LUNA Corpus is a corpus of transcribed speech, it is of a particular style: there is no sentence segmentation and punctuation. Additionally, it contains spontaneous speech artifacts such as speech disfluencies: repetitions, repairs, truncated words, etc., all of which have to be translated for a proper alignment to take place. Professional translators, on the other hand, are accustomed to working with written text. Thus, there are two translation artifacts: (1) punctuation is being inserted, which is a minor issue; and (2) speech disfluencies are *translated*, not recreated in the target language. Consequently, translation of spoken language phenomena have to be additionally inspected. Native speakers of target languages were queried for judgments on 'naturalness' of translated disfluencies and a policy was established for each language.

### 3.2. Speech Disfluency Translation Policy

The following policy was applied for speech disfluency translation. If the language pair is close enough to allow replicating disfluencies in the target language by the same morpho-syntactic means, without breaking the 'naturalness' of an utterance, they were replicated (Spanish, see example in Table 2). On the other hand, if the speech disfluency in target language requires different morpho-syntactic operation (e.g. determiner or preposition repetition in the source language is translated as a content word, postposition or suffix repetition), the disfluency is marked in text as such (Turkish, see example in Table 2). As a result, speech disfluencies are replicated in Spanish, and are marked in Turkish and Greek.

For example, in the utterance "*... ho un problema con la sta con la tastiera ... non non riesco più a usarla*" (English: 'I have a problem with the keyboard ... cannot use it anymore'), there are two speech disfluencies; and their translations are given in the Table 2. As example indicates, disfluencies are not easily replicable in every target language: e.g. for Turkish, because of word order differences and rich morphology, replication of the negation requires repetition of the whole verb, which was judged by native speakers to be 'unnatural'.

## 4. Cross-language SLU Porting

The translations of LUNA Corpus are aligned by dialog and utterance IDs; thus, the Multilingual LUNA Corpus constitutes a parallel Italian - Spanish - Turkish - Greek spoken dialog corpus readily available for translation and spoken dialog system research. Since the LUNA corpus was designed for data-driven spoken language system training, there are Italian Spoken Language Understanding modules. Multilingual LUNA Corpus was used in (Stepanov et al., 2013), and the authors present experiments on language style and domain adaptation for cross-language SLU porting. The authors compare the cross-language SLU porting between close and distant family languages (Spanish - Italian and Turkish - Italian). In this section we describe the use cases for the corpus: automatic SLU porting using Statistical Machine Translation: test-on-source (used in (Stepanov et al., 2013)) and test-on-target approaches.

### 4.1. Porting Scenarios

The two scenarios differ with respect to the direction and the object of translation. In the test-on-source scenario the direction of translation is from a language the system is being ported to (target language) to the language of the existing SLU (source language). The object of translation is user utterances in the target language. In the test-on-target scenario, on the other hand, the direction of translation is from the source language to the target language. The object of translation is the data used to train the source SLU, and new language understanding components are trained. Both scenarios have their own set of challenges; however, test-on-source scenario was repeatedly reported to yield better SLU performance (see (Stepanov et al., 2013) for references).

### 4.2. Test-on-Source

In the test-on-source scenario, the Italian Spoken Language Understanding model with the Statistical Machine Translation systems trained on Multilingual LUNA Corpus, yields concept error rate of 25.8% for Spanish and 39.2% for Turkish (see Table 3). For comparison, the original LUNA SLU has concept error rate of 21.5%.

| Scenario | ES | TR |
|---|---|---|
| *Test-on-Source* | 25.8 | 39.2 |
| *Test-on-Target: Italian references* | 29.0 | 46.5 |
| *Test-on-Target: sorted references* | 29.0 | 43.0 |

Table 3: Performance of the SLU systems ported using test-on-source and test-on-target scenarios for Spanish (ES) and Turkish (TR). For test-on-target results are for both Italian and sorted references. The results are reported as concept error rate (CER). The original Italian SLU has CER of 21.5.

### 4.3. Test-on-Target

For the test-on-target scenario, the attribute-value annotation has to be transferred from Italian to Spanish and Turkish. This could be done using different alignment approaches (see (Jabaian et al., 2013)). In line with the published works, the best results were obtained by training SLU on a corpus produced via indirect alignment. In indirect alignment a phrase alignment is used to project concepts from the source language to the target language.

Concept error rate for Spoken Language Understanding systems ported using test-on-target scenario are the following: for Spanish is 29.0% and for Turkish is 46.5% using Italian concept order as a reference (see Table 3). However, since word orders are different from language to language (especially Italian and Turkish), a relaxed evaluation by sorting concepts in alphabetical order is used. With the relaxed evaluation Spanish results remain the same, which indicates very close word order, and for Turkish they improve to 43.0%, confirming word order differences.

In line with previous research on SLU porting, test-on-source approach yields better SLU performance. The reported results indicate that porting SLU to distant family language (such as from Italian to Turkish) using either approach is challenging. Multilingual LUNA corpus is by design to support the research on this.

## 5. Conclusion

We have presented the Multilingual LUNA Corpus, a translation of Italian LUNA Corpus to Spanish, Turkish, and Greek. The corpus provides multilingual aligned data for both close and distant family languages; thus, allows for broader perspective on cross-language system portability. We described challenges and the process of manual creation of multilingual spoken conversation corpus and evaluated the created corpus on cross-language SLU porting task.

## Acknowledgments

This research is partially funded by the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 296170 – PortDial.

## 6. References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL) / the 17th International Conference on Computational Linguistics (COLING)*.

Ali Orkan Bayer and Giuseppe Riccardi. 2013. On-line adaptation of semantic models for spoken language understanding. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

Harry Bunt. 2005. A framework for dialogue act specification. In *In Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information*.

Anna Corazza, Alberto Lavelli, and Giorgio Satta. 2007. Analisi sintattica-statistica basata su costituenti. *Intelligenza Artificiale*, (2):38–39.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of Semantic Representation of Spoken Language Workshop of Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Xiaodong He, Li Deng, Dilek Hakkani-Tür, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *Proceedings of the ICASSP 2013, IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Bassam Jabaian, Laurent Besacier, and Fabrice Lefèvre. 2013. Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transaction on Audio, Speech and Language Processing*, 21(3).

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289. Morgan Kaufmann.

Geoffrey Leech and Andrew Wilson. 1996. Eagles: Recommendations for the morphosyntactic annotation of corpora.

Fabrice Lefèvre, Djamel Mostefa, Laurent Besacier, Yannick Estève, Matthieu Quignard, Nathalie Camelin, Benoit Favre, Bassam Jabaian, and Lina M. Rojas-Barahona. 2012. Leveraging study of robustness and portability of spoken language understanding systems across languages and domains: the portmedia corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Evgeny A. Stepanov, Ilya Kashkarev, Ali Orkan Bayer, Giuseppe Riccardi, and Arindam Ghosh. 2013. Language style and domain adaptation for cross-language slu porting. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*.

David Traum. 1996. Conversational agency: The trains-93 dialogue manager. In *Proceedings of Twente Workshop on Language Technology, TWLT-II*.