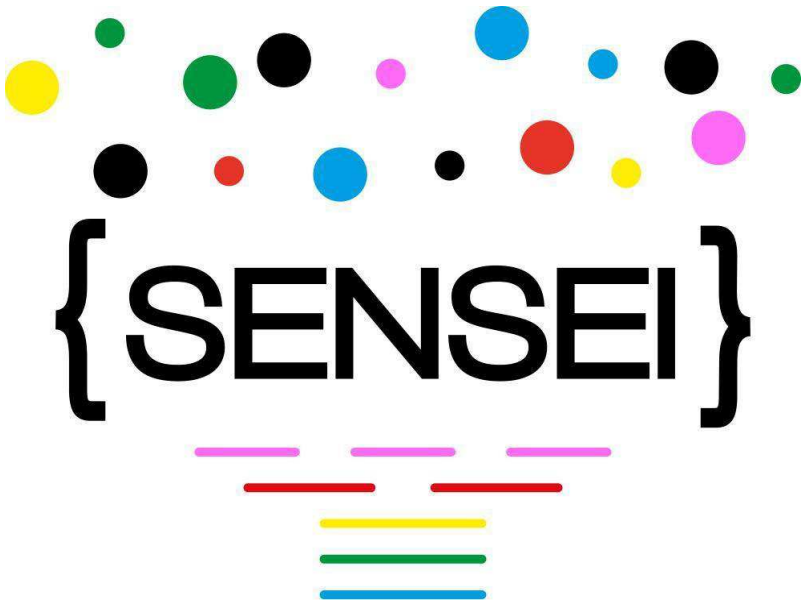


D3.1 – Report on the semantic parsing of human-human conversations (spoken and text)

Document Number	D3.1
Document Title	Report on the semantic parsing of human-human conversations (spoken and text)
Version	2.0
Status	Final
Work Package	WP3
Deliverable Type	Report
Contractual Date of Delivery	31.10.2014
Actual Date of Delivery	31.10.2014
Responsible Unit	AMU
Keyword List	FrameNet , syntactic dependency , parsing , model adaptation , cross-lingual studies
Dissemination level	PU

Editor

Frederic Bechet (AMU)

Contributors

Ahmet Aker (USFD)

Orkan Bayer (UNITN)

Benoît Favre (AMU)

Fred Bechet (AMU)

Giuseppe Riccardi (UNITN)

Evgeny Stepanov (UNITN)

Jérémy Trione (AMU)

Hugo Zaragoza (WEBSAYS)

SENSEI Coordinator

Prof. Giuseppe Riccardi

Department of Information Engineering and Computer Science

University of Trento, Italy

riccardi@disi.unitn.it

Document change record

Version	Date	Status	Author (Unit)	Description
0.1	28/07/2014	Draft	F. Bechet (AMU)	Table of Content and outline (who does what)
0.2	28/08/2014	Draft	H. Zaragoza (Websays)	Social-media document pre-processing
0.3	01/09/2014	Draft	F. Bechet (AMU)	New structure of the document. Introduction + generic architecture
0.4	02/09/2014	Draft	E. Stepanov	Change in table of content to include cross-language methodology
0.5	09/09/2014	Draft	B. Favre, F. Bechet (AMU)	Add section on syntactic parsing
0.6	15/09/2014	Draft	A. Aker (USFD)	Add description of the social media corpus used to test SEMAFOR
0.7	21/09/2014	Draft	J. Trione, F. Bechet (AMU)	Add section on the semantic annotation of the DECODA corpus
0.8	23/09/2014	Draft	A Aker (UFSD)	Add section on the target evaluation of SEMAFOR to social media data
0.9	24/09/2014	Draft	J. Trione, F. Bechet (UFSD)	Add section on the evaluation of Frame selection on ASR transcriptions
1.0	26/09/2014	Draft	O. Bayer, G. Riccardi (UNITN)	Add section 4.3 and 4.4.1
1.1	26/09/2014	Draft	A Aker (UFSD)	Add section on the frame evaluation
1.2	29/09/2014	Draft	F. Bechet (AMU)	Overview of the first draft
1.3	03/10/2014	Draft	E. Stepanov, G. Riccardi (UNITN)	Comparison cross-language vs. corpus-specific parser
1.4	03/10/2014	Draft	J. Trione (AMU)	Comparison SEMAFOR on English translation of RATP-DECODA and corpus-specific parser
1.5	06/10/2014	Draft	F. Bechet (AMU)	Executive summary + conclusion
1.6	08/10/2014	Draft	F. Bechet (AMU)	Bibliography
1.7	08/10/2014	Draft	E. Chiarani (UNITN)	Quality Check completed
1.8	13/10/2014	Draft	F. Bechet (AMU)	Last corrections before review process

1.9	15/10/2014	Draft	F. Bechet (AMU)	Corrections after review process
2.0	20/10/2014	Final	E. Chiarani, G. Riccardi	Final version

Executive summary

This report presents the models and tools developed in the first year of the SENSEI project for the semantic parsing of human-human conversations. We use in SENSEI a combination of models from flat entity-based annotation to FrameNet-based models. The frame parsers used in SENSEI are based either on off-the-shelf tools adapted on the input and output channels to the tasks and the specificities of SENSEI data; or corpus-specific parsers developed when some level of annotation is available on a given corpus on which new models can be trained.

This report describes the following studies:

- Adaptation of a syntactic parser to process spontaneous speech with limited supervision. This study showed than even with a limited supervision, an important increase in performance can be achieved by retraining parsing models with examples of the specificities of spontaneous speech compared to the “canonical” form of written language.
- Frame annotation of the French RATP-DECODA corpus. There was no large French corpus with FrameNet-based annotation available at the start of SENSEI. On of main results of WP3.1 was to develop such a corpus on the RATP-DECODA data thanks to a FrameNet model defined by the French ANR project ASFALDA and the syntactic annotation performed on RATP-DECODA.
- Evaluation of a state-of-the-art FrameNet parser (SEMAFOR - Das et al., 2010) on SENSEI data without adaptation to check the coverage and the robustness of generic semantic models to the specificities of the human-human conversation data used in SENSEI. This evaluation has been done for English on the Guardian social-media corpus and the human translation of a section of the RATP-DECODA corpus to English.
- Cross-language methodology: when no available parser or annotated corpus is available for a given language, this method consists in translating the source document to English, using a generic parser then aligning the output with the source document. This methodology is applied to the Italian LUNA data and compared to the output a corpus specific parsers that was developed during the LUNA project.

Table of Content

1. Introduction

Parsing human-human conversations consists in enriching text transcription with structural and semantic information. Such information includes sentence boundaries, syntactic and semantic parse of each sentence, para-semantic traits related to several paralinguistic dimensions (emotion, polarity, behavioral patterns) and finally discourse structure features in order to take into account the interactive nature of a conversation. Sentence or sentence-like unit segmentation as well as discourse level are in the themes of work package WP4, the other levels are in WP3.

In SENSEI, we want to take advantage of these structural and semantic cues in order to be able to obtain relevant summaries and report of human-human conversations for the different SENSEI use cases. These parsing processes are needed because coarse-grained analyses, such as keyword search, are unable to capture relevant meaning and are therefore unable to understand human dialogues.

Two kinds of “conversations” are targeted in SENSEI: spoken conversations in customer service telephone call centers and text conversations represented by messages and comments in social-media platforms and web-chat applications. These two kinds of conversations are obviously very different, however they share some very specific characteristics that justify to consider them together in SENSEI:

- non-canonical language: spontaneous speech and “WEB” language such as the one used in short messages and discussion forums are known to represent different levels of language than the “canonical' one used in written text such as newspaper articles;
- “noisy messages”: for spoken messages, automatic speech transcription systems make errors, especially when dealing with spontaneous speech; for short text messages, users do typos, spelling mistakes or use non-standard acronyms to write quickly their messages;
- relevant and superfluous information: redundancy and digression make conversation messages, both spoken and written, prone to contain superfluous information that needs to be discarded;
- conversation transcripts are not self-sufficient: for spoken messages, even with a perfect transcription, non-lexical information (prosody, voice quality) has to be added to the transcription in order to convey speakers intention (sentiment, behavior, polarity); for text messages, paralinguistic markers such as emoticons or characters format need to be processed in order to build a semantic representation of a message.

Figure 1 presents the general architecture of the WP3.1 activity: semantic parsing of human-human conversations. The general process is divided into three levels of processing: conversational data pre-processing; syntactic parsing; semantic parsing.

- Pre-processing level. This level starts with processes specific to each use-case. For the speech use-case, it involves the transcription (automatic or manual) of the spoken content and the segmentation into speakers’ turns and sentence-like units. For the social-media use case, this level starts with the removal of “noise” in WEB-data

(boilerplate detection), content extraction and text formatting.

- In WP3, the structuration of spoken or web messages is limited to the segmentation into single speaker turn or single messages. The structuration into conversations is handled in WP4.
- In addition to these fundamental processes, this level contains other Natural Language Processing (NLP) tasks such as entity extraction, sentiment detection and topic classification. These tasks are performed by generic tools as well as corpus-specific tools when annotation labels are available.
- Syntactic parsing aims to uncover the word relationships (e.g. word order, constituents) within a sentence and support the semantic layer of the language-processing pipeline. Shallow syntactic processes, including part-of-speech and syntactic chunk tagging, are usually performed in a first stage. One of the key activities described in this deliverable is the adaptation of a syntactic dependency parser to the processing of spontaneous speech. The syntactic parses obtained are used in the next step for semantic parsing.
- Semantic parsing is the process of producing semantic interpretations from words and other linguistic events that are automatically detected in a text conversation or a speech signal. Many semantic models have been proposed, ranging from formal models encoding “deep” semantic structures to shallow ones considering only the main topic of a document and its main concepts or entities. We use in SENSEI a FrameNet-based approach to semantics that, without needing a full semantic parse of a message, goes further than a simple flat translation of a message into basic concepts: FrameNet-based semantic parsers detect in a sentence the expression of frames and their roles. Because frames and roles abstract away from syntactic and lexical variation, FrameNet semantic analysis gives enhanced access to the meaning of texts: (of the kind ``who does what, and how where and when ?").

The first year of the SENSEI project for WP3 was devoted first to collect and select existing tools for processing SENSEI data according to the three levels previously described, then develop and/or adapt models when necessary to the specificities of the data collected in WP2 for the three SENSEI languages English, French and Italian

For the pre-processing steps, we used off-the-shelf tools for social media data and data-specific models for the speech use-case when annotated data was available. For syntactic parsing we used first generic parsers developed to process canonical text. Then we present a study on the adaptation of the Part-Of-Speech and dependency models of a graph-based parser to the specificities of spontaneous speech. We show that significant improvement can be obtained with a limited adaptation effort.

A similar approach was followed for semantic parsing: considering a state-of-the-art Frame-parser (SEMAFOR), we present first a study on the direct application of SEMAFOR to the SENSEI data, then we compare this approach to the development of corpus-specific frame parsers.

These studies have been made, for the speech use-case, on the French RATP-DECODA and the Italian LUNA corpus; and on the English news article+comments corpus. All this corpus are described in deliverable D2.1.

This deliverable is structured according to the pipeline displayed in Figure 1:

- Section 2 presents the pre-processing steps in conversational data, both for the speech and the social media use cases. These steps include first transcription for speech data, and cleaning for WEB data, then standard information extraction processes to detect keywords, concepts, themes, opinions, etc.
- Section 3 describes the adaptation of a dependency parser to process conversational data
- Section 4 presents the different parsing methods studied during this first year of the SENSEI project to process both speech and social media data.

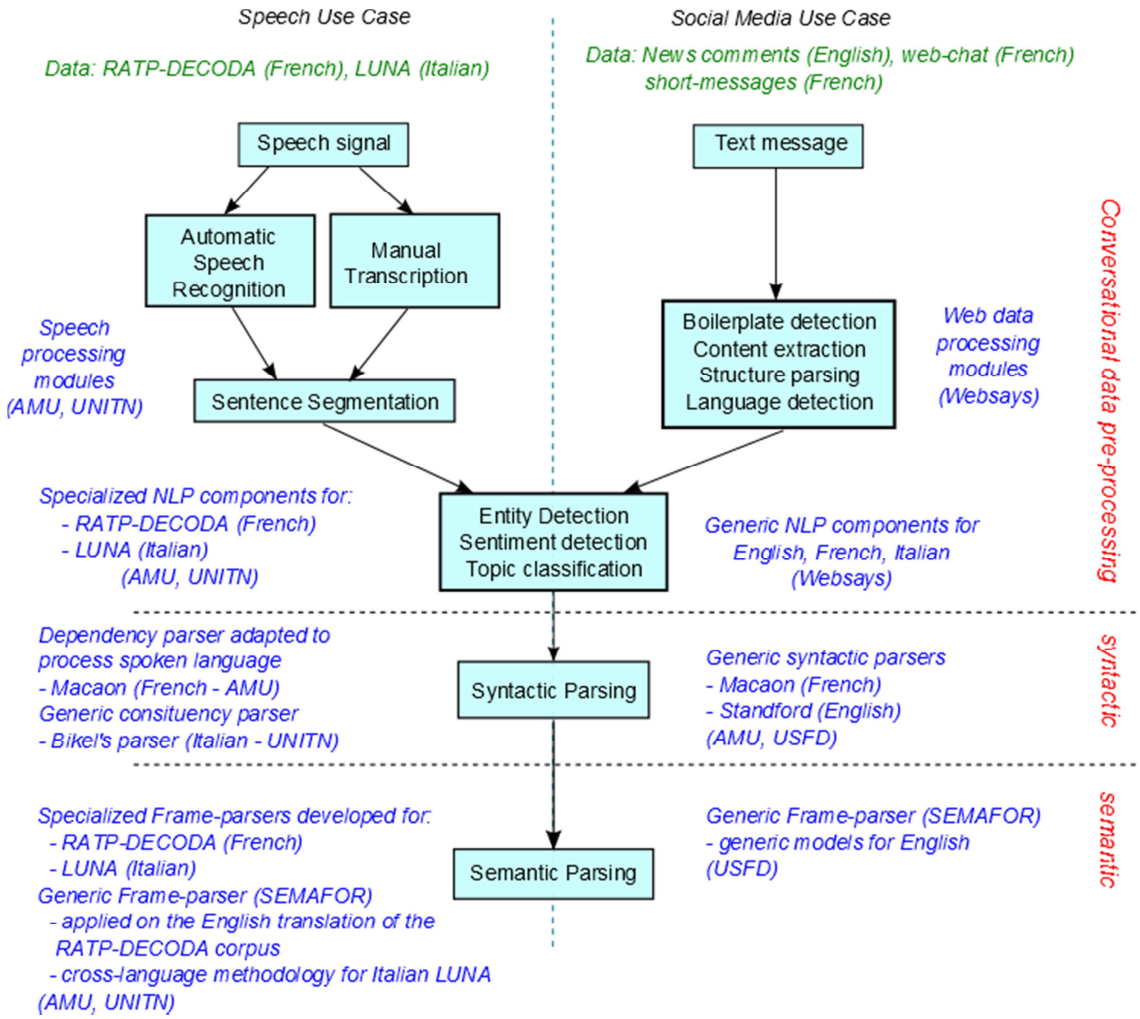


Figure 1- Description of the WP3.1 architecture: semantic parsing of human-human conversations

2. Pre-processing conversational data

2.1 Processing speech transcriptions

Telecommunication and call-centre companies involved in Customer Relationship Management (CRM), internally record call-centre conversational speech data, which are the focus of this project. Recently some European and national research programs have devoted resources to acquiring corpora through collaboration between service providers and research labs:

- In the LUNA FP6 project human-machine and human-human conversations have been collected in the context of Information Technology (IT) help-desk interactions for French and Italian.
- In France, the ANR project DECODA has recorded a large human-human spoken conversation corpus at the Paris Public Transport company (RATP) contact centre. This corpus, called RATP-DECODA, has been enriched with several level of linguistic annotations.

Both these last two corpora (LUNA for Italian, RATP-DECODA for French) are used in the experiments reported in this deliverable and are presented in deliverable D2.2.

Because these two corpora were available before the SENSEI project, we were able to directly use the human and automatic transcriptions of each of them. Human transcriptions of speech data are useful as they allow us to test the generalization capabilities of our parsing models to the specificities of spontaneous speech, without dealing with Automatic Speech Recognition issues. On the other hand, automatic transcriptions are necessary to evaluate our systems in a realistic setting. In this deliverable we will focus mainly on human transcriptions, however a first evaluation on automatic transcriptions of the RATP-DECODA data is also provided.

2.2 Processing social media data

Unlike speech data, documents crawled on the Internet from social media sources don't need any transcription and can be processed directly. However the meaningful content of these documents is often buried under superfluous data and need some levels of parsing in order to be extracted. This parsing process is made of the four following processes:

Boiler Plate Detection: Unstructured HTML content obtained by crawling (as opposed to structured content obtained by API access) is processed to remove unwanted parts (*boiler plate detection*). This is very important to remove unwanted "matches" in headers, side-bars, navigational titles and advertising.

Content Extraction: Unstructured HTML content is analyzed to detect the boundaries of relevant content and its basic metadata (*the body of the post, its title, author, date.)*

Structure Parsing: Specialized parsers are written for specific data sources in order to extract the maximum amount of information and structure. For example newspaper parsers are used to segment its pages into post, comments, comment's authors, ratings, etc. This work is detailed in the different WP2 deliverables.

Language Detection: language detection can be very challenging in short texts with brands, acronyms, URLs pieces, etc. The Websays pipeline uses a combination of methods to detect the language of a post, the main stages being:

- A fast look-up is performed for similar texts that may have been hand-labeled (i.e. a near-duplicate that has had its language label previously corrected by a human analyst), in which case the human-generated label is used. (This is extremely useful to avoid misclassifying future re-posts of posts that have been already corrected by a human analyst).
- String preprocessors remove terms that are likely to mislead the classifier (e.g. non-words, URLs, hashes, account-specific brands and acronyms, etc.)
- Unicode character heuristics are used to detect alphabet-specific languages (e.g. Japanese, Russian)
- Dictionary based frequent expressions are then used
- A character n-gram HMM is used to detect the group of most likely languages
- A topic-specific error cost-matrix is used to correct biases (or boost specific languages) for each specific topic.

2.3 Entity extraction and message-classification

Similarly to pre-processing steps, we used in SENSEI off-the-shelf and corpus-specific models that were available before the start of the SENSEI project to produce the first level of semantic annotations.

For the speech use-case, we use tools and models developed during the LUNA project (for Italian) and the DECODA project (for French). Dialog acts, concepts and Named Entities are annotated for LUNA; Named Entities and call-types are available for RATP-DECODA. From these annotations, state-of-the-art tagger and classifier are trained to produce automatic annotations (Bechet et al., 2012).

For the social media data, 4 levels of annotation are performed by the WEBSAYS pipeline:

Online-Terms Detection: a set of regular expressions are used to identity URLs, smileys, @authors, #hashes, retweet and forward notations, etc.

URL normalization: URLs in text are typically expressed as relative or partially specified paths, and they can use URL shorteners. In this step URLs are normalized and resolved so that they lead to their full unique URL.

Named Entity Detection: a combined approach is used to named entity detection:

- A dictionary-lookup method is used to detect and re-write named entities specific to the domain of the topic. These dictionaries are built on-line by human analysts directly interacting with the Websays Dashboard. After a few months of operations, topic dictionaries grow to several hundred entities and stabilize.
- A CRF model trained on a standard generic named entity corpus is used to detect named entities in English, French, Italian, Spanish and Portuguese.

Sentiment Detection: a combined approach is used to sentiment detection:

- A weighted-dictionary method is used to detect clearly positive and negative expressions. Dictionaries are structured by language and topic and can be modified directly through the Websays Dashboard by human analysts while browsing the posts. Websays dictionaries contain several thousands of terms covering six languages.
- A proprietary nearest-neighbor based method is used to detect similar posts that have been hand-labeled.

3. Syntactic Parsing

Syntactic parsing aims to uncover the word relationships (e.g. word order, constituents) within a sentence and support the semantic layer of the language-processing pipeline. Parsing is traditionally tightly connected to rewriting grammars, usually context free grammars, used together with a disambiguation model. Many current state-of-the-art text parsers are built on this model, such as (Petrov et al., 2007). Shallow syntactic processes, including part-of-speech and syntactic chunk tagging, are usually performed in the first stage.

This traditional view of parsing based on context-free grammars is not suitable for processing non-canonical text such as automatic speech transcripts or WEB media data (emails, posts, chat): due to ungrammatical structures in this kind of text, writing a generative grammar and annotating transcripts with that grammar remains difficult. New approaches to parsing based on dependency structures and discriminative machine learning techniques (McDonald et al., 2007) will be used in SENSEI for two main reasons: (a) they need less training data and (b) the annotation with syntactic dependencies of conversation transcripts is simpler than with syntactic constituents. Another advantage is that partial annotation can be performed (Bechet et al., 2009). The dependency parsing framework also generates parses much closer to meaning which eases semantic interpretation.

Using dependency parsing for speech processing has been proposed in previous studies (Bechet et al., 2009 ; Lambert et al., 2013), however the problem of the adaptation of a dependency parser to the specificities of speech transcripts, manual or automatic, of spontaneous real-world speech remains an open problem. This section describes the adaptation process of a dependency parser to spontaneous speech in order to perform open domain Spoken Language Understanding thanks to a FrameNet approach. We present why it is crucial to adapt parsers that are originally trained on written text to the specificities of spontaneous speech on manual transcriptions containing disfluencies, and discuss the usefulness of this approach to perform open-domain SLU on ASR transcriptions even with a high WER.

3.1 Spontaneous speech in the RATP-DECODA corpus

The RATP-DECODA[1] corpus consists of 1514 conversations over the phone recorded at the Paris public transport call center over a period of two days (Bechet et al., 2012). The calls are recorded for the caller and the agent, totaling over 74 hours of French-language speech.

The main problem with call-center data is that it often contains a large amount of personal data information, belonging to the clients of the call-center. The conversations collected are very difficult to anonymized, unless large amounts of signal are erased, and therefore the corpus collected can't be distributed toward the scientific community.

In the DECODA project we are dealing with the call-center of the Paris transport authority (RATP). This applicative framework is very interesting because it allows us to easily collect

[1] The RATP-DECODA corpus is available for research at the Ortolang SLDR data repository: http://sldr.org/sldr000847/fr

large amount of data, from a large range of speakers, with very few personal data. Indeed people hardly introduce themselves while phoning to obtain bus or subway directions, ask for a lost luggage or for information about the traffic. Therefore this kind of data can be anonymized without erasing a lot of signal.

While conversations last 3 minutes on average, about a third is less than one minute, 12% are longer than 5 minutes and the longest are over ten minutes. Calls usually involve only two speakers but there can be more speakers when an agent calls another service while putting the customer on wait.

Each conversation is anonymized, segmented, transcribed, annotated with disfluencies, POS tags and syntactic dependencies, topics and summaries. The call center dispenses information and customer services, and the two-day recording period covers a large range of situations such as asking for schedules, directions, fares, lost objects or administrative inquiries.

The distribution of the duration is given in the next table:

Table 1- Repartition of the calls duration

Duration	<=1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	>10
# of dialogs	597	367	230	139	68	43	27	13	10	6	14

As we can see most of dialogs are quite short, however 12% of them are longer than 5 minutes.

In the RATP-DECODA corpus, annotated disfluencies consist in repetitions, discourse markers such as hesitations, and false starts. Discourse markers are the most frequent form of disfluency, occurring in 28.2% of speech segments, repetitions occur in 8.0% of segments and false starts, the least frequent, are represented in 1.1% of segments.

Disfluencies corresponds to repetitions (e.g. le le), discourse markers (e.g. euh, bien) and false starts (e.g. bonj-). These annotations have been manually performed on the corpus thanks to a WEB-interface allowing to write regular expressions to recognize some word patterns corresponding to a kind of disfluencies. For each regular expression written, the interface displays all the matches found in the corpus in order to check the relevance of the expression. Once validated, all the occurrences matching the expression are tagged with the chosen label.

Table 2 displays the amount of disfluencies found in the corpus, according to their types, as well as the most frequent ones. As we can see, discourse markers are by far the most frequent type of disfluencies, occurring in 28% of the speech segments.

Table 2 - Distribution of the disfluencies manually annotated on the DECODA corpus

disfluency	# occ.	% of turns	10 most frequent forms
discourse markers	39125	28.2%	[euh] [hein] [ah] [ben] [voilà] [bon] [hm] [bah] [hm hm] [écoutez]
repetitions	9647	8%	[oui oui] [non non] [c' est c' est] [le le] [de de] [ouais ouais] [je je] [oui oui oui] [non non non] [ça ça]
false starts	1913	1.1%	[s-] [p-] [l-] [m-] [d-] [v-] [c-] [t-] [b-] [n-]

The DECODA corpus has been split in two subsets respectively called DEC-TRAIN (93,561 turns) and DEC-TEST (3,639 turns) that will be used for training and evaluating the syntactic parsing adaptation methods described in section 3.3.

In addition to the manual transcriptions, an Automatic Speech Recognition (ASR) process has been applied to the speech files in order to obtain realistic automatic transcriptions of conversational data. The RATP-DECODA corpus is a very challenging corpus from an ASR point of view, as many dialogues are recorded in very noisy conditions when users are calling the service in the streets, buses or metro stations.

Table 3 presents the ASR performance on the RATP-DECODA corpus we are using in this study.

Table 3 - ASR performance on the corpus, according to the kind of speaker in the call-centre

speaker	WER
caller	49.4
operator	42.4

Although the average WER is very high, not all dialogues have such poor performance. If some dialogs are too noisy to be processed, the transcriptions of most of them contain enough valid content to allow semantic processing.

3.2 Parsing models

The tagger and syntactic parser we use in this study come from the MACAON tool suite (Nasr et al., 2011). The POS-tagger is based on a linear-chain CRF as implemented in the CRFsuite

library[2]. The syntactic parser is a first-order graph-based dependency parser trained using the discriminative perceptron learning algorithm with parameter averaging (McDonald et al., 2005). Although second order parsers usually yield better results on written data, our experiments showed that first order parsers behave better on oral data. It uses the same first-order features as (Bohnet et al., 2010). Compared to transition-based parsers, graph-based parsers are particularly interesting for ASR transcriptions because they have a more even distribution of errors and are less prone to error propagation (McDonald et al., 2007). This can be explained by the fact that transition-based parsers typically use a greedy inference algorithm with rich features, whereas graph-based parsers typically use exhaustive search algorithms with limited-scope features.

We used in this study the ORFEO tagset for POS and dependency labels. The ORFEO POS tagset is made of 17 tags. Words that are part of a disfluent expression have been assigned a POS. For example, a repetition such as: ``je je je veux'' (I I I want) is tagged: ``CLI CLI CLI VRB''.

The ORFEO syntactic dependency labels tagset is restricted to 12 syntactic labels (Subject, Direct Object, Indirect Object, Modifier …) and a specific link (DISFLINK) for handling disfluencies.

The DISFLINK dependency is introduced in order to link disfluent words to the syntactic structure of the utterance. Disfluent words are systematically linked to the preceding word in the utterance. There is no deep linguistic reason for this, the only aim is to keep the tree structure of the syntactic representation. When a disfluent word starts an utterance, it is linked to a phony empty word that starts each sentence.

Table 4 - The ORFEO dependencies label tagset

SUJ	Subject
OBJ	Direct Object
OBL	Indirect Object
AUX	Auxiliary
AFF	Affix
DET	Determiner
MOD_REL	Relative Clause
MOD	Modifier

[2] http://www.chokkan.org/software/crfsuite

COORD	Coordination
DEP_COORD	Coordinated Element
ROOT	Utterance Root
DISFLINK	Disfluency

3.3 Adapting models to spontaneous speech

We describe in this section two experiments for parsing real-life spontaneous speech transcriptions as can be found in the RATP-DECODA corpus. The first one consists in simply using a parser that has been trained on written material. In the second one a speech corpus has been semi automatically annotated and a parser has been trained on it.
All experiments have been performed on DEC-TEST, which has been manually annotated using the ORFEO dependencies label tagset. This corpus is composed of 3,639 turns which correspond to 25,231 tokens. It contains 882 repetitions and 1692 occurrences of discourse markers.

The first parser was trained on the training section of the French Treebank (Abeille et al, 2003) (FTB-TRAIN). The FTB corpus is a collection of newspaper articles from the French journal Le Monde. The results are reported in the following table.

Table 5 - Parsing accuracy according to the training corpus (FTB-TRAIN or DEC-TRAIN) on the FTB-TEST and DEC-TEST corpus with and without disfluencies (for DEC-TEST)

corpus	FTB	RATP-DECODA		
train	FTB-TRAIN	FTB-TRAIN	FTB-TRAIN	DEC-TRAIN
test	FTB-TEST	DEC-TEST	DEC-TEST	DEC-TEST
		NODISF	DISF	DISF
UAS	87.92	71.01	65.78	85.90
LAS	85.54	64.28	58.28	83.86

The first column reports parsing accuracy on the FTB test set, the others on the DEC-TEST corpus from which disfluencies have been manually removed (NODISF) or kept (DISF).

Two standard metrics are used to measure the quality of the syntactic trees produced by the parser. The Unlabeled Attachment Score (UAS), which is the proportion of words in a sentence for which the right governor has been predicted by the parser and the Labeled Attachment

Score (LAS), which also takes into account the label of the dependency that links a word to its governor.

Table 5 shows that a parser trained on written material behaves poorly on spontaneous speech: the LAS drops from 85.54 to 58.28. The performances of the parser on speech from which disfluencies has been removed are intermediate, with a LAS equal to 64.28. This result is nonetheless artificial since the disfluencies have been manually removed from the parser input.

In order to adapt the parser to the specificities of oral French, we have parsed the DEC-TRAIN corpus with the parser described above and developed an iterative process consisting in manually correcting errors found in the automatic annotations thanks to a WEB-based interface (Bazillon et al. 2012). This interface allows writing regular expressions on the POS and dependency tags and the lexical forms in order to correct the annotations on the whole RATP-DECODA corpus. Then the parser is retrained with this corrected corpus. When the error rate computed on a development set is considered acceptable, this correction process stops.

The resulting corpus, although not perfect, constitutes our training corpus, obtained at a reasonably low price compared to the whole manual annotation process of the corpus.

The results of the new parser are reported in column five of Table 5.

As one can see, the accuracy of the new parser is far above the accuracy of the parser trained on the FTB even after the disfluencies have been removed. The performances of the parser can be compared to the performances of a parser for written data despite the fact that the parser has been trained on a partially manually corrected corpus.

Two reasons can explain this result. The first one is that the DECODA corpus has a quite restricted and specific vocabulary and the parser used is quite good at learning lexical affinities. The second one is that the DECODA corpus has a rather simple syntax with utterances generally restricted to simple clauses and less common ambiguities, such as prepositional attachment and coordination, than written texts.

One crucial issue is the amount of manual supervision needed to update the models. If a whole annotation of the corpus is needed, the process will be too costly whatever gain in performance is achieved.

We display in Figure 2 the learning curve of the POS tagger, starting from a generic model trained on the FTB, and including some manual annotation on the target corpus. As we can see, even a very limited annotated subset of the corpus can boost performance: by adding as little as 20 dialogs, the POS error rate drops by more than half (green curve) from 19% to 8%.

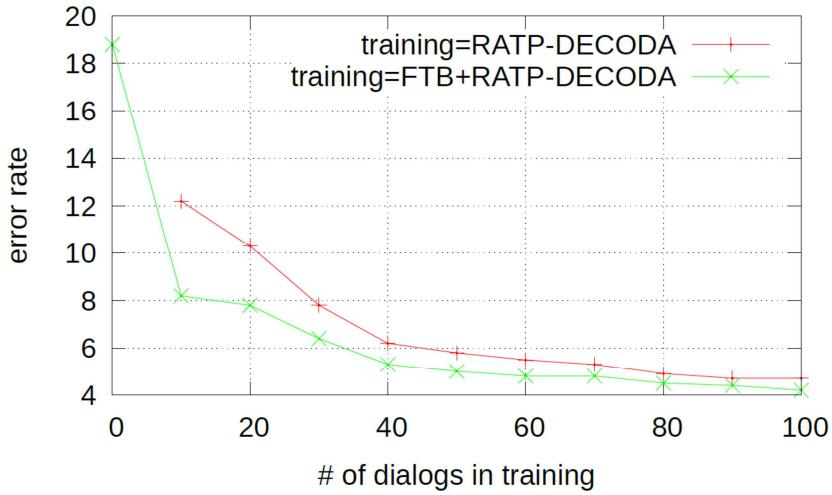


Figure 2 - Learning curve of the POS tagger with and without the FTB on the RATP-DECODA corpus

4. Semantic Parsing

Many methods have been proposed for limited domain Spoken Language Understanding (SLU), following early works on the ATIS corpus (see (Tur et al., 2011) for a review of SLU methods and models). Regardless of the paradigm chosen for performing SLU (parsing, classification, sequence labelling), the domain-ontology concepts and relations are always directly predicted from the ASR word transcriptions, sometimes with features coming from a linguistic analysis based on generic syntactic or semantic models. For open-domain SLU, it is necessary to choose an abstract level of representation that can be applied to a large range of domains and applications, therefore syntactic and semantic models developed in the Natural Language Processing community for processing text input are good candidates.

As presented in the overview, we choose a FrameNet approach to semantic in this WP. FrameNet parsing is traditionally decomposed into the following subtasks (whether applied sequentially or not):

- trigger identification: find the words that express frames. For instance in "she declared to her friend that she was going out". The target word "declared" is identified.
- trigger classification: assign the relevant frame in context (assign the frame STATEMENT to the trigger "declared")
- role filler identification: find/segment the expressions that may fill a frame role ("she", "to her friend" and "that she was going out" should be identified as potential role fillers
- role filler classification: assign the roles to the role fillers candidates ("she", "to her friend" and "that she was going out" play respectively the Speaker, Addressee and Message roles, defined for the frame STATEMENT

The last two subtasks are generally referred to as «semantic role labeling" (SRL), though this term is more general and includes SRL with other roles than that of FrameNet, in particular PropBank roles. (Gildea et al., 2002) presented the first study on role filler classification: they proposed a probabilistic classifier that, given an English sentence, a lexical trigger within that sentence and the (gold) corresponding frame, assigns FrameNet roles to syntactic phrases within the sentence. This seminal work was followed by a large number of studies, with variants using other kinds of classifiers such as maximum entropy (Fleischman et al., 2003) or SVM (Coppola et al. 2009).

Similarly to syntactic parsing, state-of-the-art semantic frame-based parsers are now available and can be trained to a new language and/or domain, as long as annotated corpora are available. We used the SEMAFOR parser in this deliverable as the state-of-the art parser. Corpora annotated with frames are available for text for English (FrameNet corpus), and partial annotation is available on speech transcription as for example the Ontonotes English Broadcast Conversation corpus annotated with verb frames. In Italian, the LUNA corpus has been annotated with the FrameNet framework. For French, no large corpus annotated with frames was available before SENSEI, therefore one of the main task of WP3.1 was to develop such a corpus.

Having spoken corpora partially or fully annotated with frames in the three SENSEI languages (English, French, and Italian), we present in this report two kinds of experiments:

- development of a corpus-specific Frame parser, when an annotated corpus is available
- use of a state-of-the art frame parser (SEMAFOR)

For the first case we used the LUNA parser developed during the LUNA project and briefly recalled here; then we present the development of the parser used to annotate the French RATP-DECODA corpus with semantic frames (Bechet et al., 2014). For the second case, two studies are reported in this deliverable:

- Using directly SEMAFOR on English data (social media corpus from The Guardian; manual English translation of the RATP-DECODA corpus);
- Using cross-language methodology to project the Italian LUNA corpus to English, then process SEMAFOR n-best hypotheses output to fit the target language/domain.

We perform a first evaluation of the performance of such approaches, and some comparison between using SEMAFOR vs. corpus-specific parsers for the speech use-cases. The cross-media and cross-domain model adaptation processes being the topic of WP3.2, we will only present in this deliverable some problems encountered when processing social-media data with models trained either on canonical written text or transcriptions of spontaneous speech. A complete evaluation of the different parsers, following the intrinsic evaluation measures presented in D1.2 will be provided at the end of Period 2.

4.1 Semi-supervised frame annotation of the RATP-DECODA corpus

We use in this study a FrameNet model adapted to French through the ASFALDA project[3]. The current model, under construction, is made of 106 frames from 9 domains. Each frame is associated to a set of Lexical Units (LU) that can trigger the occurrence of a frame in a text.

The first step, in annotating a corpus with FrameNet, is to detect LUs and generate frame hypotheses for each detection. We did this process on the RATP-DECODA corpus and found 188,231 frame hypotheses from 94 different frame definitions. We decided in this study to restrict our model to the frames generated by a verbal LU. With this filtering we obtained 146,356 frame hypotheses from 78 different frames.

Table 6 presents the top-10 frames found in our corpus. As expected the top frames are related either to the transport domain (SPACE) or the communication domain (COM and COG).

Each frame hypothesis does not necessarily correspond to a frame, most LUs are ambiguous and can trigger more than one frame or none, according to their context of occurrence.

Annotating manually with frame labels a corpus like the RATP-DECODA corpus is very costly. However we claim in this study that by merging LU detection and dependency parsing, we can produce a first frame annotation of our corpus, at a very low cost if a dependency parser is available.

[3] https://sites.google.com/site/anrasfalda

This process consists, for each verbal LU, in searching in the output of the parser for the dependencies (such as subject or object) of each selected verb. If no dependencies can be found we discard the LU. Otherwise we consider it as a frame candidate. This first annotation can be further refined by adding some semantic constraints on the possible dependent of a given LU, considering the domain of the corpus.

This process is done on the manual transcription of the spoken corpus and can be used to extract semantic patterns that can be looked for in ASR transcripts, as described in (Bechet et al. 2009).

Table 6 - Top-10 frame hypotheses in the RATP-DECODA corpus

Domain	Frame	# hyp.
SPACE	Arriving	8328
COM-LANG	Request	7174
COG-POS	FR-Awareness-Certainty-Opinion	4908
CAUSE	FR-Evidence-Explaining-the-facts	4168
COM-LANG	FR-Statement-manner-noise	3892
COM-LANG	Text-creation	3809
SPACE	Path-shape	3418
COG-POS	Becoming-aware	2338
SPACE	FR-Motion	2287
SPACE	FR-Traversing	2008

Experiments

The first experiment has been conducted on the manual transcription of the TEST corpus. This corpus has been manually annotated with POS and syntactic dependencies.

From this reference annotation we extract, in each dialogue, all verbs from the FrameNet LU lists with their dependencies. They correspond to the basic semantic structures that are needed to access to the frame level. For example, for the verb 'perdre' (to lose}) we can find the following examples in our corpus: `LOSE(I,metro-card)` in "*I have lost my my metro-card in* .."; `LOSE(daughter,teddy-bear)` in "she my daughter lost her teddy-bear in the .. ".

For example, Table 7 presents for the verb 'perdre' (to lose), some examples of these semantic structures that are found in our corpus.

Table 7 - Example of semantic structure: predicate+dependencies from the RATP-DECODA corpus

j'	ai perdu	carte
I	've lost	card
fille	a perdu	doudou
daughter	had lost	teddy bear
j'	ai perdu	journée
I	've lost	day
j'	ai perdu	travail
I	've lost	job

These dependency structures are the target of our evaluation: we measure how well we can detect them with an automatic parser instead of manual reference annotations.

We compare in Table 8 the performance of the two parsers presented in Section 3, the one trained only on the FTB and the one adapted to the RATP-DECODA corpus.

Average Precision and Recall in the detection of LUs with dependencies are presented in Table 8. As we can see, the performance of the adapted parser has a much higher precision than the standard models.

Table 8 - Performance detection of semantic dependency structures on the manual transcriptions of the RATP-DECODA corpus

parser	precision	recall	f-measure
FTB	75.9	85.5	77.3
TRAIN	88.2	88.4	87.2

The second experiment has been conducted on the Automatic Speech Recognition (ASR) transcription of the corpus. The RATP-DECODA corpus is a very challenging corpus from an ASR point of view, as many dialogues are recorded in very noisy conditions when users are calling the service in the streets, buses or metro stations.

Although the average WER is very high, not all dialogues have such poor performance. Table 9 shows the performance in terms of average Precision and Recall in the detection of LUs with dependencies in the ASR transcriptions. Two conditions are compared: ***dep1*** compares full predicate+dependency recovery, ***dep2*** accepts partial match on the dependencies. The performance are rather limited, however considering the high WER of the transcriptions, the adapted dependency models show some robustness in these difficult conditions.

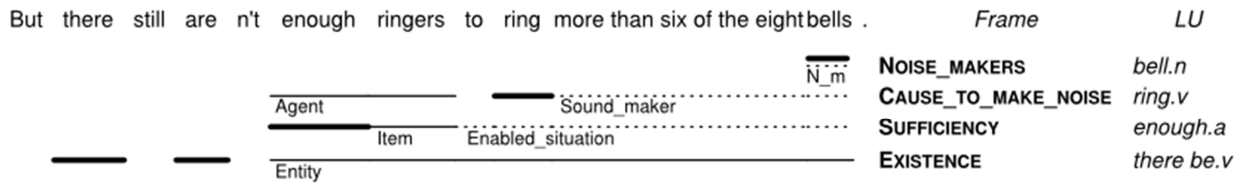
Table 9 - Performance detection of semantic dependency structures on the ASR transcriptions

condition	precision	recall	f-measure
dep1	47.4	66.4	51.4
dep2	57.3	80.3	62.7

4.2 Using SEMAFOR as a generic frame parser

SEMAFOR is described in the paper "*Probabilistic Frame-Semantic Parsing*" (Das et al. 2010). This parser is based on the theory of frame semantics[4] and statistical machine learning to produce shallow semantic structures from raw natural language text.

An example of parse, provided by the SEMAFOR developer group[5], is shown below:



**from https://code.google.com/p/semafor-semantic-parser*

Figure 3 – Example of FrameNet parsing with SEMAFOR

Each row under the sentence corresponds to a semantic frame and its set of corresponding arguments. Thick lines indicate targets that evoke frames; thin solid/dotted lines with labels indicate arguments. *N_m* under "bells" is short for the *Noise_maker* role of the *NOISE_MAKERS* frame.

The SEMAFOR parser as available for English, and uses the FrameNet 1.5 lexicon as a reference for analyzing text. Even with annotated data, the adaptation of SEMAFOR to a new language is not straightforward as it relies on external semantic resources such as WordNet, and the code of some module is language-dependent.

[4] http://framenet.icsi.berkeley.edu

[5] https://code.google.com/p/semafor-semantic-parser This parser has been developed by Dipanjan Das, Andre Martins, Nathan Schneider, Desai Chen and Noah A. Smith at Carnegie Mellon University.

Recently, in (Chen et al., 2013), SEMAFOR was used to process spontaneous speech for the development of a Spoken Dialog System. No adaptation to the specificities of spontaneous speech was performed on the linguistic models of the parser: the authors used directly the parser trained on written text, the adaptation was done on the output of the parser.

We followed this strategy on 2 use-cases: the Guardian social-media data for English; and the human translation into English of a section of the French RATP-DECODA corpus.

4.2.1Social Media use case

In this section we report semantic parsing for the social media data and evaluate the performance of the SEMAFOR parser on this data.

Data used

We run SEMAFOR on social news data collected automatically from The Guardian[6]. Statistics about the articles and comments are shown in the table below.

Table 10 - News data statistics

Type	Size	Min-sent.	Max-sent.	Avg. sent.	Min-tokens per sent.	Max-tokens per sent.	Avg.-tokens per sent.
Articles	339	11	310	46.3	151	5734	21.97
Comments	14410	2	55	4.6	9	856	13.9

Semafor output

The output from Semafor contains three classes of units:

1. a list of targets
2. frames for each target
3. frame elements (a list of arguments with their frames)

E.g. for the following sentence: ***You seriously think bad guys are created by the west?***

The output of Semafor is:


```
{"frames":[
{"target":{"name":"Importance","spans":[{"start":1,"end":2,"text":"seriously"}]},"anno
tationSets":[{"rank":0,"score":39.65277850019617,
"frameElements":[
{"name":"Interested_party","spans":[{"start":0,"end":1,"text":"You"}]}]}]},
{"target":{"name":"Opinion","spans":[{"start":2,"end":3,"text":"think"}]},
"annotationSets":[{"rank":0,"score":40.60358421917837,
"frameElements":[{"name":"Cognizer","spans":[{"start":0,"end":1,"text":"You"}]},
{"name":"Opinion","spans":[{"start":3,"end":10,"text":"bad guys
are created by the west"}]}]}]},
{"target":{"name":"Desirability","spans":[{"start":3,"end":4,"text":"bad"}]},
"annotationSets":[{"rank":0,"score":38.734898861844066,
```


[6] For this analysis the dump 20140714_BigSampleData_TheGuardian was used.

```
"frameElements":[{"name":"Evaluee","spans":[{"start":4,"end":5,"text":"guys"}]}]}]},{"
target":{"name":"People","spans":[{"start":4,"end":5,"text":"guys"}]},
"annotationSets":[{"rank":0,"score":32.35141094025542,
"frameElements":[{"name":"Person","spans":[{"start":4,"end":5,"text":"guys"}]}]}]},
{"target":{"name":"Intentionally_create","spans":[{"start":6,"end":7,"text":
"created"}]},"annotationSets":[{"rank":0,"score":46.61311479940858,
"frameElements":[{"name":"Created_entity","spans":[{"start":3,"end":5,"text":"bad
guys"}]}, {"name":"Time","spans":[{"start":7,"end":10,"text":"by the
west"}]}]}]},{"target":{"name":"Locative_relation","spans":[{"start":9,"end":10,"text"
:"west"}]}, "annotationSets":[{"rank":0,"score":22.846515985692918,
"frameElements":[{"name":"Figure","spans":[{"start":3,"end":5,"text":"bad
guys"}]}]}]}],"tokens":["You","seriously","think","bad","guys","are","created","by","t
he","west","?"]}
```


Evaluation

In this report we evaluate two of these three output classes: the targets and the frames. We perform the evaluation on sentences taken from the news articles as well as sentences selected from comments provided by the news readers. For each data type (i.e. news articles and readers' comments) we have selected random 50 sentences from the data summarized in Table 10. These sentences have been manually annotated to provide golden standard data for subsequent Semafor output evaluation. For each of the 100 sentences from this evaluation data set a human evaluator has extracted targets, i.e. words that evoke frames, using as an annotation guideline descriptions and examples in Ruppenhofer et al. (2010), in particular those related to full-text annotation. In a second step the FrameNet was searched for the frames for each target, manually disambiguating the word senses where necessary. Both the steps were performed independently of the Semafor output, i.e. the parser output was not corrected, but rather a separate human generated data set was created.

Evaluation of "targets" Lexical Units

The evaluation of targets (or Lexical Units) should provide us with insight whether Semafor correctly identifies words that evoke semantic frames in the news article and comment data and also what coverage it achieves in target identification. This is achieved by comparing the human annotated targets with the ones identified by Semafor. The results of the evaluation are shown in Table 11.

Table 11 - Target evaluation results

Type	Recall	Precision	F1-Score
Article sentences	0.6	0.75	0.66
Comment sentences	0.62	0.74	0.64

From the results in Table 11 we can see that the Semafor tool obtains 75% precision for the article sentences and 74% for the comment sentences. Between the two types of texts there is only a slight difference (1%). We think this is due to the nature of comments. Guardian comments are generally written in well-formed sentences with few to no special characters (e.g.

emoticons). In this sense they are rather comparable to the news articles. However, they do occasionally contain misspellings and a few lexemes typically not found in news articles (e.g. spoken language equivalents like "*yeah right*", or British English jargon, e.g. "*lass*").

Both precision results are somewhat lower with the precision scores obtained for the SemEval 2007[7] data, which was constructed for the purpose of semantic parsing evaluation. Semafor was trained and tested on this data and achieved 89.92% precision in target identification (Das et al. 2010). The difference of 14-15% in precision shows that Semafor loses some quality when moved from the "known world" to a new one but nevertheless achieves satisfactory results. However, it would be still better to train it on the news domain to achieve similar results as reported for the SemEval data.

The recall figures are lower than the precision scores. They are also again lower to the ones reported on the SemEval 2007 data. For this data it was reported that Semafor obtains 70.79% in recall. Again this shows that Semafor would benefit from adaptation to a new domain.

Evaluation of "frames"

For evaluating the frames we took the targets from the Semafor outputs which were correct according to the manual annotation, i.e. the target outputs of Semafor which were also identified by the human annotator. Note for both article sentence and comment targets we randomly selected 250 matching targets for the evaluation. The human annotator manually checked whether FrameNet contains a correct frame definition for each of these 500 targets. If yes, the name of the frame definition was recorded. If, on the other hand, the frame for a potential target was missing entirely, or if a target's particular word sense could not be found in FrameNet, the target was marked as having no frame definition. In the evaluation we compare the frame definition Semafor identified with the ones human has found in FrameNet. We report only precision scores as recall will be the same, i.e. we ask for each matching target whether the identified frame definition is according to the human annotator correct or not. Figures in Table 12 highlight the results.

Table 12 - Target-frame definition evaluation

Type	Precision
Comment Targets	0.56
Article Sentence Targets	0.40

From Table 12 we can see that frame definition identified for the comment sentences is substantially higher than for the article sentences. For comment sentences we have a precision of 56% and for the article sentences 40%. Reported precision figure for the SemEval07 data is

[7] http://nlp.cs.swarthmore.edu/semeval/

69.75%. This again shows that there is a gap between the “known” and “unknown” environments and this could be closed with adaptions of Semafor to the news domain.

4.2.2 Speech use case

One of the outcomes of the WP7 on dissemination will be the organization of a shared-task in an international evaluation program “Multiling 2015”.

To this purpose, and with the objective of increasing the number of participants, we decided to translate to English some sections of the Italian and French corpora of call-centre conversations.

For task WP3.1, we used a section of 50 dialogs, manually translated to English, to check the robustness of SEMAFOR applied to speech data.

Table 13 - Section of the RATP-DECODA corpus manually translated to English and used for the SEMAFOR experiments

# dialogs	# speaker turns	# words (French)	# words (English)
50	3902	25501	24479

As for the social media use case, we applied directly SEMAFOR to each dialog turn, then collect the Frame hypotheses. We were able to compare the hypotheses produced to those obtained during the annotation of the French RATP-DECODA corpus presented in section 7.1.

Since the annotation of this corpus was done in a semi-supervised way, we don’t have human frame annotation on which we could directly evaluate SEMAFOR output. We are in the process of building such a reference, in a similar way as the social media use case presented in the previous section. However prior to this qualitative evaluation, we report here a quantitative evaluation about the comparison between the annotations of SEMAFOR and those of the Frencg RATP-DECODA corpus.

On the original French version of these 50 dialogs we had, in addition to the Frames representing Named Entities, the expression of 9 frames: *Motion, Arriving, Locating, Awareness, Request, Self_Motion, Path_shape, Ride_vehicule*. These frames were triggered by 2545 lexical units.

On the English corpus SEMAFOR produced 302 different frames triggered by 7422 lexical units.

This very large difference between the number of frames output by SEMAFOR and those annotated in the French corpus has three main reasons:

1. Firstly the RATP-DECODA frame parser only focuses on frames relevant to the semantic domain of the corpus. Being corpus-specific, the annotation doesn’t need to cover all possible situations in the corpus, but rather focuses on the application domain (Paris transport system) and the conversation or behavioral domain. SEMAFOR on the other hand, being a generic tool, has a larger coverage.

2. Secondly the mismatch between the written language on which SEMAFOR was trained, and the spoken language used in RATP-DECODA leads to increase ambiguity and has a tendency to over-generate Frames. For example the lexical unit "*past*" triggers the frame "*Individual_history*" in the sentence "*it's quarter past*".

3. Lastly a lot of frames correspond to phenomenon annotated at other levels of the specific parsing process of the RATP-DECODA corpus such as Named-Entities (frames *Quantity, Cardinal_numbers, Roadways*, …) or disfluencies (frame *Sounds*).

The following table present the 20 most frequent frames output by SEMAFOR.

Table 14 - 20 most frequent frames output by SEMAFOR on the English section of the RATP-DECODA corpus

Frame	#
Quantity	348
Being_obligated	307
Cardinal_numbers	279
Locative_relation	265
Intentionally_act	261
Calendric_unit	225
Existence	211
Arriving	185
Motion	161
Temporal_collocation	159
Statement	156
Becoming	151
Capability	149
Causation	148

Judgment_direct_address	148
Roadways	146
Vehicle	142
Request	124
Placing	108
Awareness	101

The following examples show some problems due to the ambiguities of spoken language and the over-generation of frames.

Table 15 - Example of over-generation of frames by SEMAFOR on the RATP-DECODA corpus

I get it you are leaving from Menilmontant.
#Frame: Arriving (get) #FE: theme (I)
#Frame: Giving (leaving) #FE: Theme (You), recipient (from Menilmontant)
Please hold on a minute Madam
#Frame: Containing (hold) #FE: Container (Please), Contents (on a minute Madam)
Yeah yeah I get it yeah
#Frame: Arriving (get) #FE: Theme (I), Goal (it)
And we will phone back
#Frame: Contacting (phone)
#Frame: Observable_body_part (back)

Despite these problems of over generation, most of the frames annotated in RATP-DECODA can be found in the SEMAFOR output, as presented in the table 16.

Expect for the 2 frames *Path_shape* and *Ride_vehicule* which are specific to the transport domain and annotated differently between the two corpora, we can see that for most frames the SEMAFOR output contains more detection than those of the French corpus. This validates the methodology of using SEMAFOR in a first step then having a corpus-specific decision module that can filter and correct the generic hypotheses produced.

Table 16 - Intersection between the frames annotated in the French RATP-DECODA corpus and those annotated by SEMAFOR (on the 50 dialogs manually translated to English)

Frame	En	Fr
Motion	161	295
Arriving	185	47
Locating	21	21
Awareness	101	181
Request	124	32
Self_Motion	19	5
Path_shape	0	27
Ride_vehicule	0	106

4.3 Using SEMAFOR through cross-language methodology

4.3.1Methodology

Building an automatic semantic parser system for low resource languages suffers from the small amount of annotated data. Cross-language methodology solves this problem by projecting the problem from the low resource source language to a rich resource target language. SEMAFOR is the state-of-the-art semantic parser for English, therefore the semantic parsing problem for low resource languages like Italian can be transferred to the target language, English, by using the following methodology.

The cross-language methodology consists of the following steps: (1) Statistical machine translation (SMT) from the source language to the target language. (2) Semantic parsing in the target language by using the state-of-the-art semantic parser, SEMAFOR. (3) Transferring the semantic interpretations to the source language by using phrase alignments (that are extracted at the SMT step). (4) Re-scoring of multiple hypotheses by an in-domain semantic model on the source language.

Two different approaches can be used for SMT. In the first approach an off-the-shelf translation system can be used. On the other hand, a SMT system can be trained by using a parallel in-domain data by using an open-source system. The major advantage of off-the-shelf systems is the, it allows users to obtain satisfactory translations without the any expertise on SMT system training.

However, these systems are general domain and they are trained on written text, which results in performance drop when domain specific data with a different style, like speech transcription, are used. As shown in Stepanov et al., 2013; the performance of these systems can be improved by performing language style adaptation. Language style adaptation constitutes the following steps for speech transcriptions: (1) Automatic punctuation insertion, automatic case restoration, and de-tokenization. The baseline system uses *Google translate* as the off-the-shelf SMT system without the language style adaptation step.

After the SMT is carried out, semantic parsing can be performed on the output, which is in the target language (English). For transferring as many semantic interpretations as possible to the source language, SEMAFOR semantic parser is modified to output scores and n-best hypotheses for frame identification and frame element identification. Therefore, after the output of the SMT system is semantically parsed and n-best hypotheses are generated. The next step transfers the semantic interpretations in the target language to the source language by using phrase alignments obtained from the SMT system. The final step re-scores these semantic interpretations by using a in-domain semantic model. The baseline system considers the transfer of the best hypothesis without performing re-scoring.

4.3.2Comparison with the LUNA corpus-specific parser

The semantic parsing for LUNA in Italian is performed by using the following architecture that is described in Coppola et al., 2008. The architecture has 4 components. The first one is to detect

frame-evoking words, target detection. This is performed by using a rule-based approach. The second component performs frame disambiguation, this step assigns a proper frame to the target word. The semantic parser outputs every possible frame with the probability of that frame in the in-domain data. The main focus of this system is on the last two components, boundary detection and role classification. Boundary detection assigns spans to all the required frame elements (with respect to FrameNet definition). Boundary detection is modeled as a binary classification problem over the nodes of the syntactic parse tree. The role classification component assigns semantic roles to the spans detected at the previous step. This component models the problem as a multi-class classification over the syntactic parser tree nodes. The classification for the last two components is performed by using support vector machines using a combination of polynomial and tree kernels. The semantic parser for LUNA depends on the output of a constituency parser. The performance of this system is given for boundary detection (BD) and role classification (RC) with boundary detection in Table 17.

A web application for the semantic parser is available at:

http://cicerone.disi.unitn.it/FrameSemantics/Demo.php

Table 17 - The performance of the Italian semantic parser on LUNA

	Precision	Recall	F1-Score
BD	0.89	0.86	0.87
BD+RC	0.76	0.74	0.75

Experiments

The in-domain and in-language LUNA FrameNet parser (Coppola et al., 2009) and out-of-domain SEMAFOR parser using cross-language methodology are compared in terms of Frame Recognition performance, i.e. whether a presence of an individual frame is recognized in an utterance, disregarding the Frame Elements.

For example, whether a parser recognizes the frame “Using” being present in the utterance. The evaluation is done in terms of precision, recall and f-measure. For both parsers only the 1st best hypothesis is taken into consideration, i.e. no hypotheses re-ranking.

Frame Recognition evaluation does not require the full cross-language pipeline. It is sufficient to translate the source language utterances to English, which in our case was performed using Google Translate. Consequently, the reported results are the baseline performance of the cross-language approach.

The input to the LUNA semantic parser and Google Translate is identical. However, it underwent minor pre-processing, which consists of character normalization and ‘period insertion’ steps.

In the former step Italian orthographic variants of the accented characters were normalized to their canonical forms. For example, ***e`*** and ***e'*** are normalized to ***è***. In the latter step a sentence final period is inserted. For both semantic parsers – LUNA and SEMAFOR – features extracted from automatic parses are used for classification.

Due to the fact that LUNA Corpus contains corpus-specific frames, for the fair evaluation of cross-language approach with SEMAFOR we report performances with and without considering these corpus-specific frames.

The Test Set for LUNA FrameNet annotation consists of 20 dialogs (1,146 turns) that contain 1,038 fames (145 unique).

After removing the corpus-specific frames we are left with 958 frames (142 unique).

Table 18 - Comparison of LUNA in-domain semantic parser and SEMAFOR using cross-language approach performances on Frame Recognition.

	P	R	F1
LUNA (LUNA)	0.40	0.59	0.48
LUNA (FrameNet)	0.39	0.58	0.47
SEMAFOR (LUNA)	0.27	0.28	0.27
SEMAFOR (FrameNet)	0.27	0.31	0.29

**The frame sets are given in parentheses: LUNA -- all frames annotated in LUNA Corpus (i.e. including corpus-specific frames); FrameNet -- excluding corpus-specific frames.*

Even though removing corpus specific frames improves the performance of the SEMAFOR, the difference is still almost 0.20.

However, the gap in performance between two parsers is expected, since LUNA is trained on in-domain data and does not suffer from translation errors. However, LUNA frame vocabulary is significantly smaller (174 + 20 corpus-specific (Dinarelli et al., 2009)); consequently, the range of applications is limited in comparison to SEMAFOR, which considers 877 frames from FrameNet 1.5 Release (Das et al., 2014).

The future direction of the cross-language methodology is the domain and genre adaptation for both the SMT and the SEMAFOR parser.

5. Conclusions

We have presented in this deliverable the semantic models and the parsing methodology developed in WP3 for processing the Human-Human SENSEI conversations for social-media and speech data. At the end of Y1 we have studied different solutions for producing these semantic representations using either generic or corpus-specific models and tools for the tree SENSEI languages (English, French, and Italian).

If corpus-specific parsers can be considered reliable if enough effort in the corpus annotation is made, the use of generic models with or without cross-language methodology produces, as expected, lower performance than those reported in the state-of-the-art on documents similar to those used to train the models.

We will use the output of all these parsers in the summarization tasks of WP5 to check the usefulness of such a representation, even with a certain level of noise.

The next step that will be presented in the next deliverable D3.2 at the end of Period 2 is dedicated to the cross-media and cross-domain model adaptation through the answer to the two following questions:

1. How can we process and adapt the output of the generic parser SEMAFOR to improve the coverage and the precision in the frame parsing process when dealing with social-media and speech data?
2. How can we project the high precision models developed for a specific media and a specific domain to a new media and/or domain with a limited human supervision?

A complete evaluation of the different parsers that will be developed in WP3, following the intrinsic evaluation measures presented in D1.2 will be provided at the end of Period 2.

References

A. Abeillé, L. Clément, and F. Toussenel, “Building a treebank for french,” in Treebanks, A. Abeillé, Ed. Dordrecht: Kluwer, 2003.

(Bazillon et al., 2012) Thierry Bazillon, Melanie Deplano, Frédéric Béchet, Alexis Nasr, Benoît Favre: Syntactic annotation of spontaneous speech: application to call-center conversation data. LREC 2012: 1338-1342

(Bechet et al., 2014) F Bechet, A Nasr, B Favre, “Adapting dependency parsing to spontaneous speech for open domain spoken language understanding”, Fifteenth Annual Conference of ISCA, INTERSPEECH 2014

(Bechet et al., 2012) Frédéric Béchet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Bèze, Renato De Mori, Eric Arbillot “DECODA: a call-centre human-human spoken conversation corpus”. LREC 2012: 1343-1347

(Bechet et al., 2009) F. Bechet, A Nasr, "Robust dependency parsing for Spoken Language Understanding of spontaneous speech ," Interspeech 2009

(Bohnet 2010) Bohnet, Bernd, “Top Accuracy and Fast Dependency Parsing is not a Contradiction, Proceedings of COLING 2010.

(Chen et al., 2013) Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, “Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing,” in Automatic Speech Recognition and Understanding (ASRU), IEEE, 2013, pp. 120–125.

(Coppola, 2009) B. Coppola, A. Moschitti, and G. Riccardi, “Shallow semantic parsing for spoken language understanding,” in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion, Volume: Short Papers, Association for Computational Linguistics, 2009, pp. 85–88.

(Das et al. 2010) Das, Dipanjan and Schneider, Nathan and Chen, Desai and Smith, Noah A, “Probabilistic frame-semantic parsing”, Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, pages 948--956

(Das et al., 2014) D. Das, D. Chen, AFT Martins, N Schneider, NA Smith. “Frame-semantic parsing”. Computational Linguistics 40 (1), 9-56. 2014.

(Dinarelli et al., 2009) “Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics”

(Gildea et al., 2002) Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” Comput. Linguist., vol. 28, no. 3, pp. 245–288, 2002

(Fleischman et al., 2003) M. Fleischman, N. Kwon, and E. Hovy, “Maximum entropy models for FrameNet classification,” in Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, 2003, pp. 49–56.

(Hahn et al., 2011) Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato De Mori, Hermann Ney, and Giuseppe Riccardi "Comparing Stochastic

Approaches to Spoken Language Understanding in Multiple Languages," IEEE Transactions On Audio, Speech, And Language Processing, 2011.

(Lambert et al., 2013) B. Lambert, B. Raj, and R. Singh, “Discrimintively trained dependency language modeling for conversational speech recognition,” in Proc. Interspeech ISCA, 2013.

(McDonald et al., 2005) R. McDonald, K. Crammer, F. Peirera, "Online large-margin training of dependency parsers," ACL 2005.

(McDonald et al, 2007) R. T. McDonald and J. Nivre, “Characterizing the errors of data-driven dependency parsing models.” In EMNLP-CoNLL, 2007, pp. 122–131

(Nasr et al. 2011) A. Nasr, F. Bechet, J. Rey, B. Favre, and J. Le Roux, “Macaon: An NLP tool suite for processing word lattices,” Proceedings of the ACL 2011 System Demonstration, pp. 86–91, 2011

(Pado et al., 2005) S. Pado, M. Lapata, "Cross-lingual bootstrapping of semantic lexicons: The case of framenet ," AAAI 2005.

(Petrov et al., 2007) S. Petrov, D. Klein, "Learning and Inference for Hierarchically Split PCFGs ," AAAI 2007.

(Ruppenhofer et al.) Ruppenhofer, Josef, et al. FrameNet II: Extended Theory and Practice. Berkeley, California: International Computer Science Institute, 2010.

(Stepanov et al., 2013) Stepanov E. A., Kashkarev I., Bayer A. O., Riccardi G., Ghosh A., “Language Style and Domain Adaptation for Cross-Language SLU Porting” in Proc. IEEE Workshop Automatic Speech Recognition Understanding (ASRU 2013), Olomouc, Czech Republic, pp. 144-149, 2013.

(Tur et al., 2011) Gokhan Tur and Dilek Hakkani-Tür Human/Human Conversation Understanding Chapter 9 of (Tur & De Mori, 2011)

(Wang et al., 2011) Ye-Yi Wang, Li Deng and Alex Acero, "Semantic Frame Based Spoken Language Understanding," Chapter 3 of (Tur & De Mori, 2011)