{SENSEI}

D2.2 – Data Collection Report Y1

Document Number	D2.2
Document Title	Data Collection Report Y1
Version	3.2
Status	Final
Work Package	WP2
Deliverable Type	Report
Contractual Date of Delivery	31.10.2014
Actual Date of Delivery	31.10.2014
Responsible Unit	Teleperformance
Keyword List	Data Collection
Dissemination level	PU

Editor

Vincenzo Lanzolla (TP)

Contributors

Ahmet Aker	(USFD)
Emma Barker	(USFD)
Fabio Celli	(UNITN)
Benoit Favre	(AMU)
Adam Funk	(USFD)
Rob Gaizauskas	(USFD)
Vincenzo Lanzolla	(TP)
Marçal Juan Llaó	(Websays)
Marco Martinez	(Websays)
Evgeny Stepanov	(UNITN)
Veronique Tolsan	(Websays)
Jorge Valderrama	(Websays)
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
1.0	29/08/2014	Draft	Hugo Zaragoza(Websays)	Table of Content
1.1	5/09/2014	Draft	Benoit Favre (AMU)	Section 2.1 drafted
1.2	7/09/2014	Draft	Marco Martinez (Websays)	Section 3. drafted
1.3	7/09/2014	Draft	Marçal Juan, Vernoique Tolsan, Hugo Zaragoza (Websays)	Section 3. extended
1.4	8/09/2014	Draft	Jorge Valderrama (Websasy)	Section3.2 Data Schema
1.5	08/09/2014	Draft	Vincenzo Lanzolla (TP) Hugo Zaragoza (Websays)	Section 1 drafted
1.6	22/09/2014	Draft	Vincenzo Lanzolla (TP)	Section 2.3 added
1.7	23/09/2014	Draft	Fabio Celli (UNITN)	3.2.1 data schema
1.8	23/09/2014	Draft	Evgeny Stepanov (UNITN)	Section 2.2 drafted
1.9	24/09/2014	Draft	Giuseppe Riccardi(UNITN)	Revision of the doc and feedbacks
1.10	25/09/2014	Draft	Hugo Zaragoza (Websays)	Improvement of Sections 1 and 3 as per comments
2.0	26/09/2014	Draft	Letizia Molinari(TP)	More about data
2.1	26/09/2014	Draft	Vincenzo Lanzolla (TP)	Corrections and rewrites as per remarks
2.2	1/10/2014	Draft	Vincenzo Lanzolla (TP)	Formatting and more about data
2.3	06/10/2014	Draft	Elisa Chiarani (UNITN)	Quality Check
2.4	08/10/2014	Draft	Vincenzo Lanzolla(TP)	All section. Final version ready for scientific review
2.5	09/10/2014	Draft	Benoit Favre (AMU)	Scientific Review
2.6	10/10/2014	Draft	Vincenzo Lanzolla(TP)	Appendix C added. Changes as per remarks
2.7	17/10/2014	Draft	Fred Bechet (AMU)	Scientific Review
2.8	17/10/2014	Draft	Hugo Zaragoza (Websays), Fabio Celli	Addressing review's comments
3.0	20/10/2014	Final	Elisa Chiarani, Giuseppe Riccardi (UNITN)	Finalisation of the deliverable for Reviewers
3.1	28/10/2014	Final	A. Aker (USFD), V. Lanzolla (TP)	Section 3.10 added.
3.2	30/10/2014	Final	E. Barker, R. Gaizauskas	Revised section 3.10. Final

			(USFD)	version

Executive summary

Deliverable 2.2 provides specifications of data requirements for the entire projects, including the nature of data, the method of collection planned, the selected sources for each media and language.

The document reports information about data publication and sharing beyond the consortium, and the methods to obtain copy-righted free materials.

The deliverable describes the data collected during the first year of the project, the call center annotation efforts and developed tools to annotate the selected set of conversation in Italian and French language.

Finally the document describes the first collection of web data coming from social media channels, multimedia content sites and the work carried out for content extraction, pre-processing and indexing of web data.

Table of Content

1. Overview

The D2.2 speech data collection is composed of data files of speech and social media, annotated with low-level structure and pre-processed. The data collection is composed of several sub-collections:

- Speech
  - DECODA
  - LUNA
- Social Media
  - General News Topics
  - NewsPaper Publications
  - RATP
  - Orange

A number of intermediate tasks were necessary to achieve this deliverable, and are briefly described in this document:

- Study of use cases to extend data sources in order to obtain appropriate data for experimentation
- Adaptation of Data Schema and tools for the SENSEI AOF (Agent Observation Form) and segment turn selection
- Definition of a Social Data Schema
- Adaptation of the Websays parsers to the required sources
- Development of new Websays crawling facilities for online demand of sources
- Evaluation and validation of the obtained data

The different sub-collections and the tasks carried out to prepare them are discussed in the different sections of this document.

D2.2 data collection is a continuation of the preliminary collection presented in D2.1. For this reason part of the discussion was already presented (in a rougher form) in D2.1. We have chosen to present here a comprehensive picture of all data collection details up to day, copying material from D2.1 (while correcting and extending it) when necessary, without making continuous references to the previous work.

The main work carried out since D2.1 is itemized here:

- Speech Data
  - Review of TP annotation work on the DECODA corpus
  - Definition of the correct monitoring form to use for listening LUNA and DECODA conversation, based on AMU guidelines

- Implementation of the SENSEI Web Annotation
- Annotation Agent oriented summarizes
- Annotation of conversation caller/requester-oriented (synopsis) on SENSEI Web Annotation
- Internal (TP) and external (UNITN) Calibration

- Social Data
  - New data sources requiring specialized parsers were added to the pool
  - Existing topics for data collection were manually evaluated and extended
  - New topics for data collection for developed to cover the use cases
  - The asynchronous crawling infrastructure had to be extended to meet new specifications
  - Manual evaluation of collected data revealed many inconsistencies and errors that had to be diagnosed and corrected.
  - The schema had to be revised and slightly modified to introduce new meta-data

1.2 Approach

The ultimate goal of WP2 is to provide a unified data view of “conversations”, both from speech dialogues and online (typed) dialogues. This however requires a high level of abstraction from the raw data, which is not readily available; indeed, building such an abstraction is one of the main objectives of the SENSEI project.

WP2 should provide views on the data in a way that the full original data could be reconstructed. Additional annotation on the data should be provided by other WPs in the form of stand-off annotations on these views. A mapping between data schemas should be developed to achieve a unified conversation schema starting in this deliverable and completing this work in D2.3.

1.3 Data Access

1.3.1 Public Data Access

The initial data set contains three parts and over 1M items. A small sample of all the collections are provided for public online access from the SENSEI web site, together with this document, which provides an overview of the data and instructions about how to request the entire data sets. The method of data acquisition and usage is discussed in D8.2 – Ethical Issues Plan. Here we provide a summary, mainly repeating the same information, recalling the most relevant information fully contained in deliverable D8.2.

For the Social Media collection, the website provides a data bundle for D2.1: a small sample of 1000 social media items from the Social Media collection, together with the entire list of public URLs of the data crawled for this collection. The entire collection (as well as individual parts of the collection) can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it.

For LUNA data we provide a small complete sample; the entire collection is distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).

For DECODA data we provide a small complete sample. The entire collection is distributed by SLDR/Ortolang (http://crdo.up.univ-aix.fr, ID: http://sldr.org/sldr000847). Researchers or practitioners may get access to the annotated corpus of human conversations free of charge by accepting the SLDR/ORTOLANG license.

For the Teleperformance data (limited to the annotations produced by QA Supervisors during the filling of AOFs), is available to the partners internally since D2.1 and D2.2 constitutes the first public installment of the data. Similar to the social media data, the Teleperformance data can be made available to the public upon e-mail request to sensei-data@list.disi.unitn.it.

1.3.2 Partner's Data Access

For partners, a SVN data repository has been setup on one of the SENSEI servers containing all the data for easy access. In the case of the LUNA collection, the data will be distributed as-is to partners for evaluation and annotation through the data sharing agreement provided in the Ethical Issues Plan (D8.2).

The Websays Dashboard has also been made available to all partners in order to provide a rich visual interface to browse the Social Media portion of the data.

All partners have web access, upon authentication, to the **SENSEI ACOF Annotation** developed for Sensei, where they can find the DECODA and LUNA conversations, the Agent Observation Forms and synopsis registered by TP Quality Assurance.

2. Speech

2.1 Decoda Collection

The RATP-DECODA corpus consists of 1500 speech conversations recorded at the RATP call center in France. These recordings of French speaking customers and agents have been collected during the ANR DECODA project (Bechet, et al. 2012). Each conversation is available in anonymized speech, manual transcript and various layers of annotations such as sentence boundaries, part-of-speech tags, chunks, syntactic dependencies, topic boundaries, named entities, disfluencies, noises and metadata. Those conversations have been recorded over the course of a single day from a public transportation call center. The topics covered range from passenger routing, general information, complaints, etc. Table 1 describes the most frequent topics.

Table 1: Top 10 topics in the Decoda corpus

Topic	**%**
Informations	22.5
Route planning	17.2
Lost and found	15.9
Registration card	11.4
Timetable	4.6
Ticket	4.5
Specialized calls	4.5
Empty	3.6
New registration	3.4
Price info	3.0

A balanced subset of 200 conversations has been selected for further manual annotation in the SENSEI project (AOF, synopsis, semantic frames). This subset follows the same topic distribution as the whole corpus. Conversation duration ranges from 55 seconds to 16 minutes. The corpus contains 82k words, 13k sentences, an average of 414 words and 66 sentences per conversation.

The 200 conversations have been annotated with at least two synopses. The synopses are short summaries of what happens within a conversation. In a first annotation round, the relevant material included, the length and the style of the summaries was left at the annotators' discretion. The average length observed is between 6% and 7% of the number of words of the original conversations. After this round of synopsis writing, an annotation guide was produced in order to ensure the consistency of future collection and collection by other partners. This guide is used by TP for the annotation of synopses on the different corpora.

Examples of synopses:

Annotator 1:

- What bus for Gare de Lyon to Montparnasse.
- RER E timetable from Meaux to Gare de l'Est.

Annotator 2:

- Query for a bus line to go from Gare de Lyon to Gare Montparnasse.
- Query for the train timetable from Gare de Maux to Gare de l'Est at a given time.

In order to complement the syntactic annotation of the corpus, a semi-supervised full-text semantic frame annotation process was developed on conversation transcripts and will be also applied to synopsis. This process is described in depth in deliverable D3.1.The 200 dialogs and synopses have been translated to English in order to use them in the course of a shared task. Translating speech transcripts is not an easy task for professional translators as the style is informal and it is crucial for the success of the shared task that the transcripts remain faithful to the original, especially in term of disfluencies and speech artifacts. For 50% of the data, we used professional translators who had been specifically trained for the task and whose work has been validated by a quality assurance process. For the remaining 50% of the data, we have used automatic translation with the Moses system trained on the first half of the data. This gives realistic contrastive conditions for the shared task. More information on the shared task can be found in D7.2.

2.2 Luna Collection

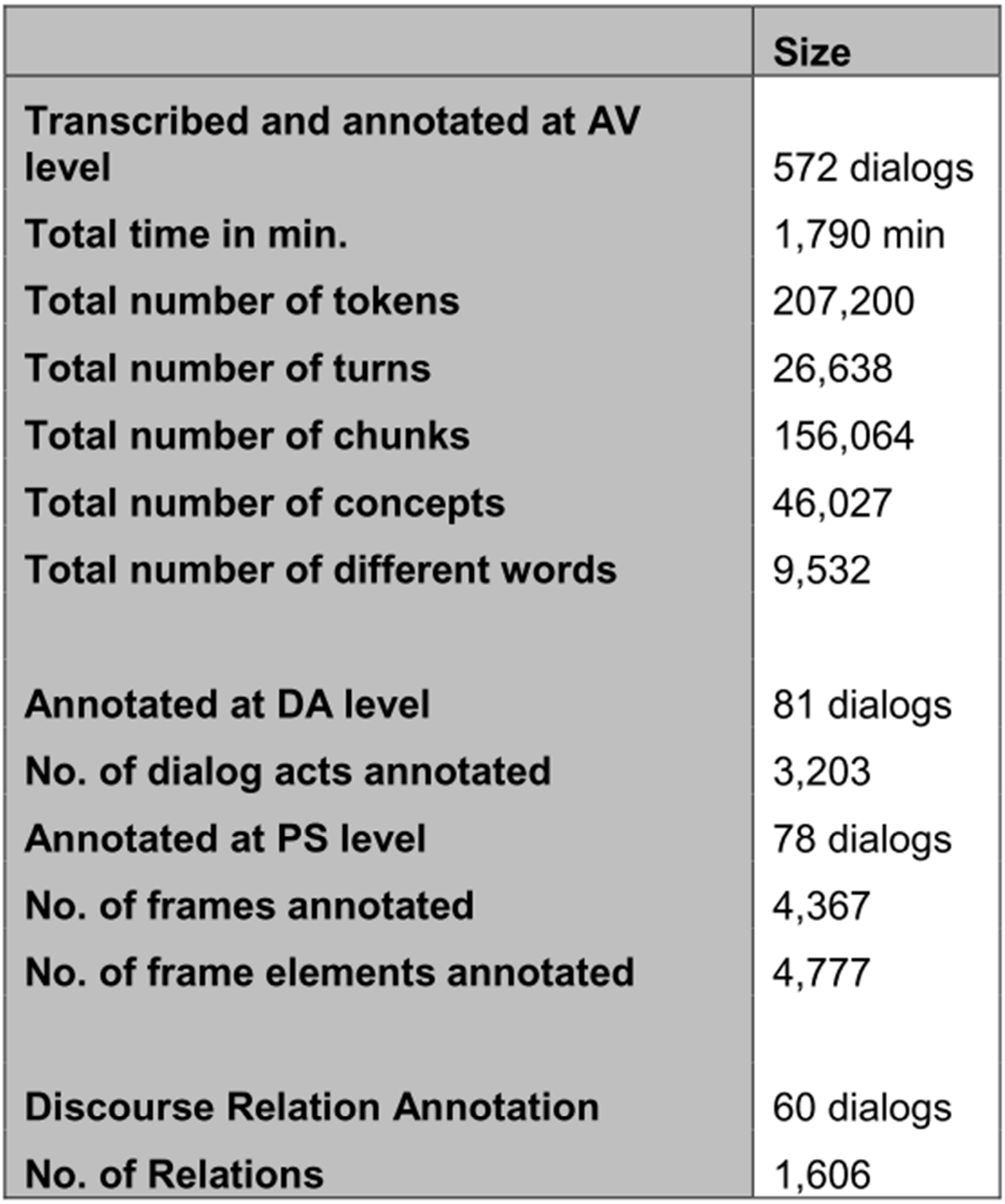
The Italian LUNA Corpus is a collection of 572 human-human dialogs in the hardware/software help desk domain. The dialogs are conversations of the users and operators involved in problem solving. The dialogs are organized in transcriptions and annotations defined within FP6 LUNA Project. The dialogs 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 the EAGLES corpora annotation standard. Attribute-value annotation makes use of a predefined ontology of domain concepts and their relations. Predicate argument annotation is based on the FrameNet model. Dialog act annotation was inspired by DAMSL, TRAINS and DIT++ and is used to mark intentions in an utterance. Discourse relation annotation was performed following the Penn Discourse Treebank (PDTB) approach.

The general process of annotation can be seen in the figure 1 below. 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 achieved semi-automatically using the Bikel parser trained on an Italian corpus with subsequent manual correction.

Table 2 below provides general statistics on the LUNA Corpus, such as the number of dialogs annotated at each level, as well as token and turn counts.

Table 2: general statistics on the LUNA Corpus



	Size
Transcribed and annotated at AV level	572 dialogs
Total time in min.	1,790 min
Total number of tokens	207,200
Total number of turns	26,638
Total number of chunks	156,064
Total number of concepts	46,027
Total number of different words	9,532
Annotated at DA level	81 dialogs
No. of dialog acts annotated	3,203
Annotated at PS level	78 dialogs
No. of frames annotated	4,367
No. of frame elements annotated	4,777
Discourse Relation Annotation	60 dialogs
No. of Relations	1,606

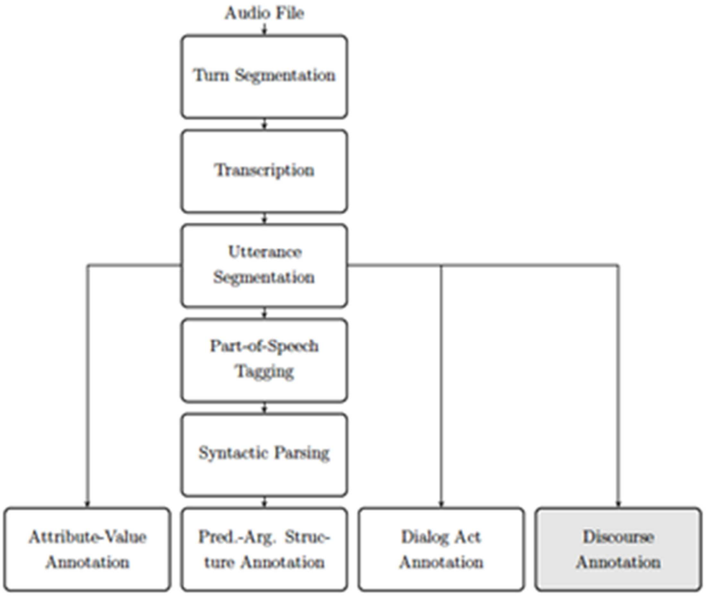


Figure 1: LUNA Annotation process

Additionally, the **LUNA Human-Human Corpus** has gone through an additional quality control/cleaning procedure.

1. Attribute-value annotation (concept ontology) is normalized;
2. Due to the nature of the hardware/software Help Desk domain, the corpus contains words borrowed from English. Thus, the corpus is corrected for misspelled words via a semi-automatic procedure (automatic detection of misspelled words with human decision on the correct form);
3. Since the LUNA Corpus contains sensitive private information, such as personal names, phone numbers, etc., which is protected by Italian privacy laws, the corpus is anonymized. A special attention is given to preserve the distribution of token within anonymized concept values. However, transcriptions and predicate-argument structure layers are not anonymized due to different segmentation and tokenization, which makes it difficult to distinguish sensitive vs. non-sensitive data (see the Figure above on the annotation process).

Within the SENSEI project a subset of 200 dialogs was selected for annotation with AOF and synopsis summaries. The criterion for dialog selection was that they contain the most levels of annotations – attribute value, dialog act, predicate argument structure, and discourse relation. Each dialog was annotated with a long and a short synopsis (summary) by different annotators.

A subset of 100 of these dialogs was selected for manual translation to English using professional translation services. Since speech transcriptions are rich in artifacts such as disfluencies and fillers, as well as lack punctuation information; and professional translators are not accustomed to dealing with such material; a methodology and a translation manual was iteratively developed. The remaining 100 dialogs of the subset will be translated using automatic machine translation.

2.3 Teleperformance data annotation

Teleperformance Quality Assurance professionals have annotated a set of LUNA and DECODA conversations, according to the requirements agreed with other WPs.

To support Quality Assurance professionals in the annotation task, a web user interface tool, named SENSEI ACOF Annotation, has been developed. With the monitoring view of the tool, described below in the document, Quality Assurance professionals can fill the items of the AOF, select the segment turn that are relevant for the evaluation of each item of AOF, and finally fill the synopsis (COF).

For LUNA (audio recordings in Italian language) 821 Agent Observation Form have been annotated for 200 distinct dialogs. For each dialog, an AOF is filled (without associating answers to speech turns in the recordings), and two synopsis are created. This annotation is called COF (Conversation Oriented Form).

Each conversation has been listened and evaluated an average of 4-5 times from different evaluators (QA professionals).

In total, 1642 synopses were collected.

The average qualitative score annotated is 69%, this mean that communications skills of the agents are not high-level.

Table 3: Statistics for LUNA

QA professionals	Number of AOF	Number of COF	Score Evaluation Weighted Average
Annotator1	200	400	72
Annotator2	202	404	72
Annotator3	205	410	70
Annotator4	34	68	54
Annotator5	180	360	61
TOTAL	**821**	**1642**	**69**

For DECODA (audio recordings in French language) 95 Agent Observation Forms have been collected in excel format, because the SENSEI AOF annotation tool was not ready in June and July for this collection.

Starting from September 2014 we have moved the content of DECODA AOF into the SENSEI web tool, adding turn references, and COF, and we have generated a total of 222 AOF for 118 distinct conversations.

The total number of collected synopses is 444.

The average qualitative score generated is 84%, this mean that communications skills of the agents are medium-level.

Table 4: Statistics for DECODA

QA professionals	Number of AOF	Number of COF	Score Evaluation Weighted Average
Annotator1	108	216	84
Annotator2	114	228	84
TOTAL	**222**	**444**	**84**

The TP Quality Assurance Team, in October, has integrated the 200 LUNA conversations with the segment turn data, using a new feature of the SENSEI AOF annotation tool, that allow users to drag&drop the relevant segment turn for each item.

By the end of October, TP concluded the evaluation and annotation work, producing the complete AOF, COF and segment turns for 200 distinct LUNA dialogs and 118 DECODA conversations.

In Appendix B is presented the current state of data collection and annotation.

2.3.1 The SENSEI ACOF Annotation tool

2.3.1.1 Overview

One of the objectives of the SENSEI project is to produce Monitoring forms (AOF), synopses (COF) and relevant segment turns for a selected set of DECODA and LUNA data.

The SENSEI ACOF Annotation tool provides Quality Assurance supervisors with a user friendly web interface to fill the SENSEI AOF and the synopsis for each conversation, saving the data in a relational database for reporting purposes and future use.

The quality Assurance professionals, with the monitoring view can fill the items of the AOF, select the segment turn that are relevant for the evaluation of each item using drag&drop feature and fill the synopsis. The segment turns are saved in the database and can be used by other WPs for training automatic prototypes, or by QA professionals for quickly locating problematic speech in the conversations.

Starting listen conversation, the QA professional follow specific item of the Agent Observation form and select relevant segment turn. During that phases QA professional evaluate how the agent manage the call and requests coming from end users, if the agent is able to give correct information and answer, if the agent understand and resolve the problem if it is possible, in a right way and using courtesy and professionalism.

Some example of item are:

- Advisor listens actively and asks relevant questions?
- Advisor shows the information in a clear, comprehensive and essential way?
- Advisor uses positive words?

During the second part of the monitoring view (COF) QA professional generate synopsis in order to clarify the reason of the call.

The Sensei ACOF tool has been developed following the specifications described in Appendix B and Appendix C of D1.2.

2.3.1.2 Objectives

The objectives of the Sensei ACOF Annotation tool are two folds. First, to improve the productivity and accuracy of quality assurance professionals. Second, to build a collection of annotated data that are saved in a database in the format agreed on with other WP leaders.

The Quality Assurance professionals can carry out more tasks at once: they can listen to the audio, see the transcription and fill the AOF and the synopsis in a unique page at the same time.

To have data saved in a structured way in a database, allow the generation of reports and the comparison with other automatically generated annotations.

2.3.1.3 Main Features

The access to the tool requires authentication.

Users can change the language of the user interface by selecting a value from a menu-list of three available languages (Italian, English and French)

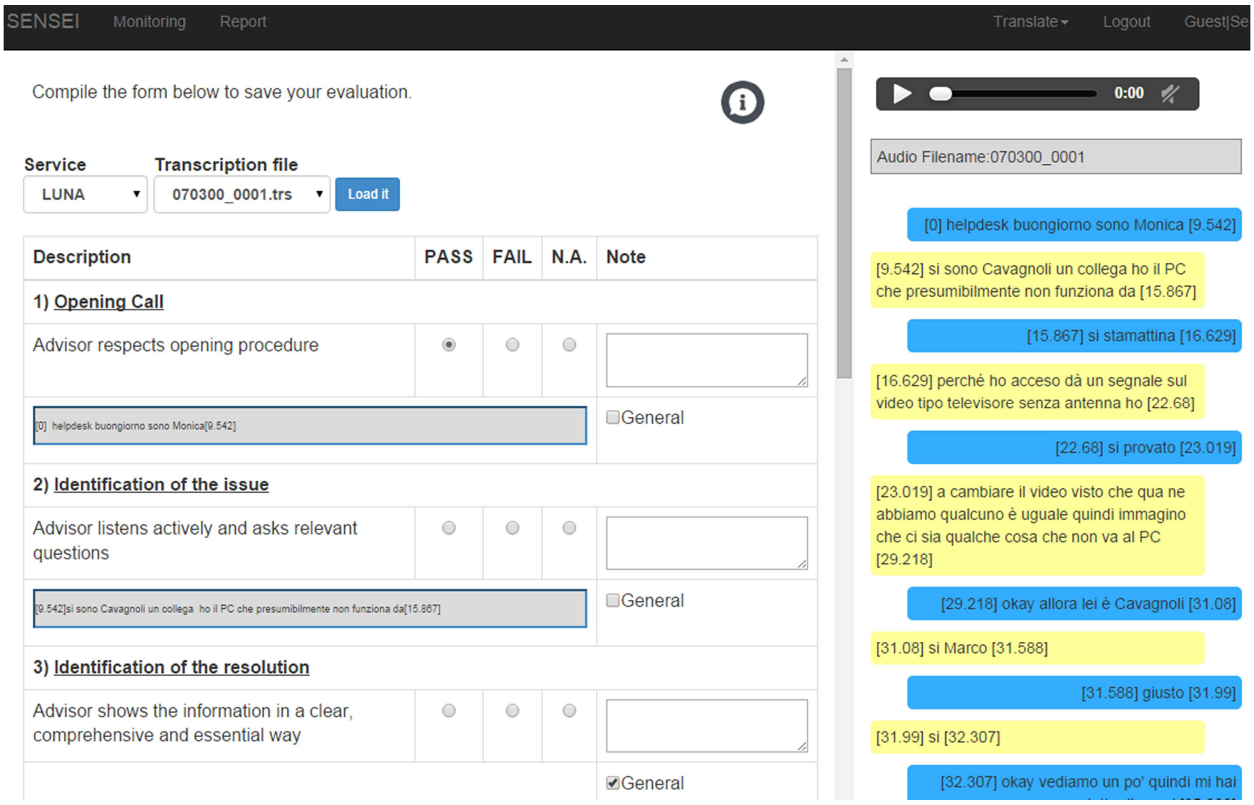
The tool has two main views, Monitoring and Report.

2.3.1.4 Monitoring view

The Monitoring view allows the user to perform the following tasks: select the domain between LUNA and DECODA, select the transcript file, listen to the audio of the call, see the conversation's transcription, fill SENSEI Agent Observations Forms using drag&drop feature to select the relevant speech turns, write the synopses.

The help on line shows the guidelines for filling the synopsis.

Figure 2 below is a snapshot of the monitoring view and shows its main features.



Compile the form below to save your evaluation.

Service: LUNA

Transcription file: 070300_0001.trs

Description	PASS	FAIL	N.A.	Note
1) Opening Call				
Advisor respects opening procedure	●	○	○	
[0] helpdesk buongiorno sono Monica[9.542]				□General
2) Identification of the issue				
Advisor listens actively and asks relevant questions	○	○	○	
[9.542]si sono Cavagnoli un collega ho il PC che presumibilmente non funziona da[15.867]				□General
3) Identification of the resolution				
Advisor shows the information in a clear, comprehensive and essential way	○	○	○	
				☑General

Audio Filename:070300_0001

[0] helpdesk buongiorno sono Monica [9.542]

[9.542] si sono Cavagnoli un collega ho il PC che presumibilmente non funziona da [15.867]

[15.867] si stamattina [16.629]

[16.629] perché ho acceso dà un segnale sul video tipo televisore senza antenna ho [22.68]

[22.68] si provato [23.019]

[23.019] a cambiare il video visto che qua ne abbiamo qualcuno è uguale quindi immagino che ci sia qualche cosa che non va al PC [29.218]

[29.218] okay allora lei è Cavagnoli [31.08]

[31.08] si Marco [31.588]

[31.588] giusto [31.99]

[31.99] si [32.307]

[32.307] okay vediamo un po' quindi mi hai

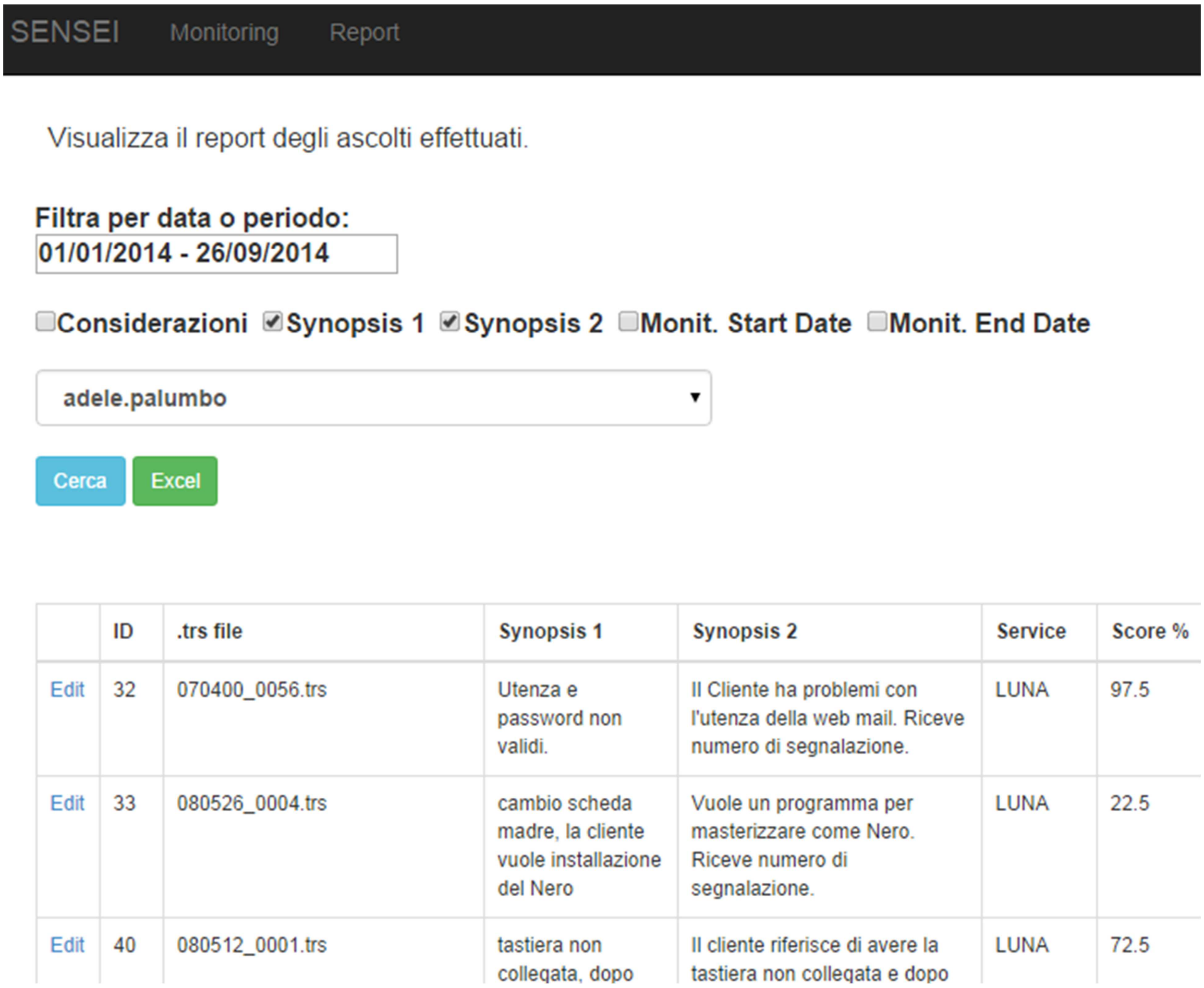
Figure 2: screenshot of the Monitoring view.

2.3.1.5 Report view

The Report view lets users extract records that match filtering conditions. It’s possible to download the result as an excel spreadsheet.

The available filtering conditions are the monitoring’s date and the user who registered the monitoring form.

Figure 3 below is a snapshot of the Report view.



Visualizza il report degli ascolti effettuati.

Filtra per data o periodo: 01/01/2014 - 26/09/2014

□ Considerazioni ☑ Synopsis 1 ☑ Synopsis 2 □ Monit. Start Date □ Monit. End Date

adele.palumbo ▾

	ID	.trs file	Synopsis 1	Synopsis 2	Service	Score %
Edit	32	070400_0056.trs	Utenza e password non validi.	Il Cliente ha problemi con l'utenza della web mail. Riceve numero di segnalazione.	LUNA	97.5
Edit	33	080526_0004.trs	cambio scheda madre, la cliente vuole installazione del Nero	Vuole un programma per masterizzare come Nero. Riceve numero di segnalazione.	LUNA	22.5
Edit	40	080512_0001.trs	tastiera non collegata, dopo	Il cliente riferisce di avere la tastiera non collegata e dopo	LUNA	72.5

Figure 3: snapshot of the Report view

3. Social Media

3.1 Previous Work in D2.1

Starting in month 1, partners have worked towards defining a rich data schema for the collection of data and metadata from social media. We considered the structure of many different social media, including blogs, Twitter, Facebook and Youtube, and specially newspaper forums which contain the most complex dialogue structure (with comments to comments, voting on posts and comments, etc.)

Delivery D2.1 already proposed an extensive data schema which allowed us to disseminate early versions of data dumps and begin work. This data schema was based extensively on previous data schemas developed by Websays for the storage of structured social media data, and was extended to accommodate many meta-data items and linked references present in news forums.

After D2.1 work began analyzing data dumps in different ways. Websays analysts manually check results from many sources against the original data sources to validate the results, account for missing metadata, etc. This way, a number of bugs were found in the parsers (and their Unit Tests) and were remediated. In parallel different partners started using the data for their own investigations, to find inconsistencies, partial data, etc. This was again analyzed to remediate the errors and re-crawl and re-parse the affected sources. In some cases this required fine-tuning the schema to add or modify fields.

3.2 Data Schema

Since the specifications for the prototype and demonstrators are not yet definitive, we brainstormed around “reasonable” data uses and derived from these the data fields (data specifications) required. This exercise was based mostly in existing data uses from the research community, Teleperformance and Websays.

The unit of publication is defined as the minimum unit “posted” by an author, typically smaller than a web-page or post-view since these may contain many comments and other forms of multiple-author interactions. We denote a “clipping” (or “post”) this unit of publication, and define the schema around this unit. A clipping represents for example a blog post (such as a blog post, a Facebook post, a Tweet or Retweet) and a different clipping will represent a comment associated to with a post, a retweet, etc.

Clippings are represented in a way that the entire conversation can be reconstructed afterwards: they are indexed by author and the various available author_IDs (e.g. the Facebook apiAuthorID), date of publication, position in the comment thread, etc. They maintain pointers to their parent post (e.g. in the case of a comment). Furthermore they contain additional meta-data (e.g. the number of likes).

Two types of pointers are recorded: postID and parentID. The postID pointer allows for fast retrieval of elements within a post (without the need of recursive calls). The ordered-tree structure of comments to posts and comments to comments is preserved by the parentID and the position pointers. This is illustrated in the Figure 4 below:

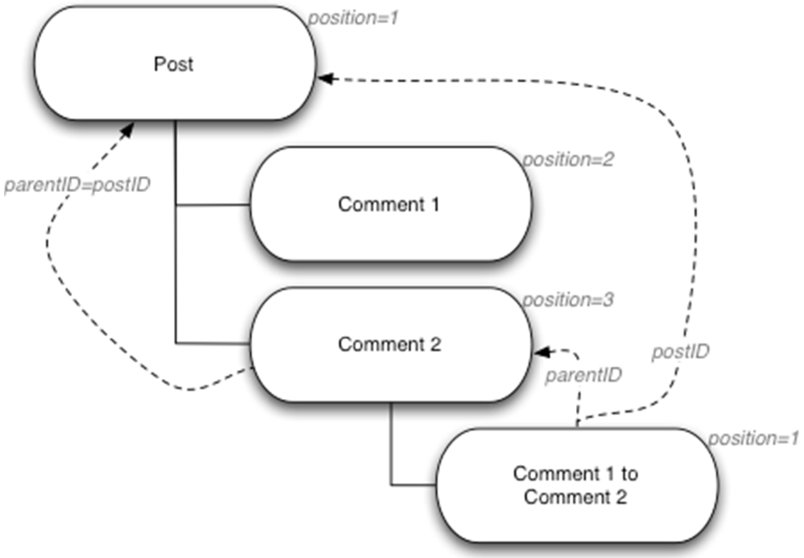


Figure 4: Illustration of the relationship between parentID, postID and position.

The resulting schema (the D2.2 data schema) is presented in Appendix A and specifies over 50 data fields per post, the main categories being:

- Post IDs (such as postID and parentID, URL, domain, etc.)
- Pos Type (such as article/comment, social/forum, etc.)
- Post Data (such as title)
- Post and Author Metadata (such as number of comments, number of followers)
- Preprocessor Annotations (such as sentiment)
- Timestamps, Localization, Metrics and NLP annotations

This schema maps into an XML representation of every post. The XML Schema is also presented in Appendix A.

3.2.1 CorEA Corpus Data schema

Beside the general data, retrieved by Websays following the schema above, UniTN collected a corpus, named CorEA, for training and testing parasemantic information extraction and for visualization. CorEA data has been retrieved from Corriere.it that provides a lot of metadata about participants to a conversation. We defined a specific data schema for it, reported in Table 5 below:

Table 5: CorEA Corpus Data schema

Data Type	Data Field Description
IDs	message Id participant Id participant's nickname
Metadata	data type (e.g. article/comment) text timestamp macro topic category comment reference to parent participant comment reference to parent comment link to participant's picture count of replies to the comment count of likes of the message participant's activity score count of interests of participant participant page views count of messages of participant count of shares count of participant's votes indignation score disappointment score preoccupation score amusement score satisfaction score
Annotation	agreement/disagreement labels

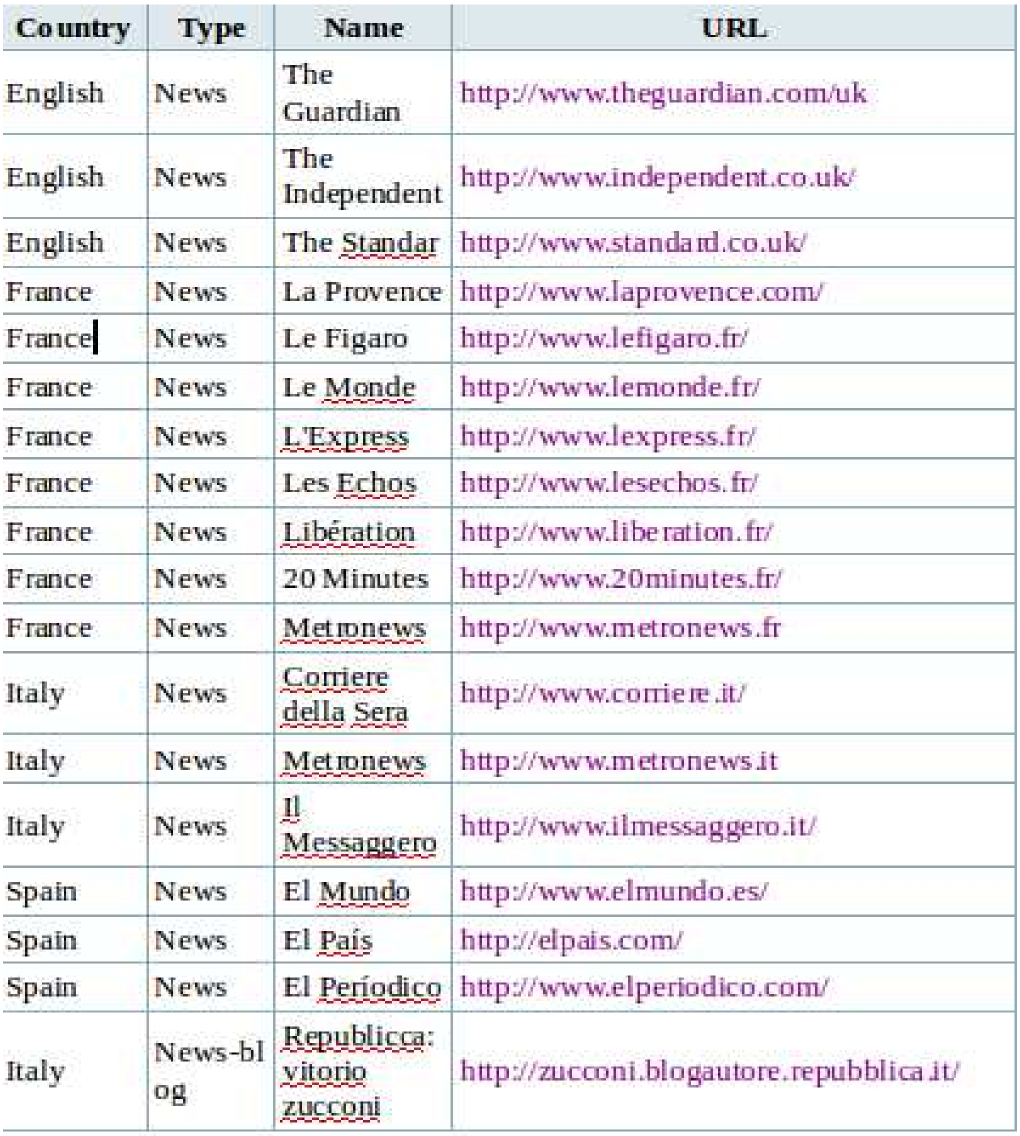
This data is being made available under this data schema at D2.2 for preliminary research, and if it proves interesting it will be integrated with the remaining corpus under the general schema in D2.3.

3.3 Data Sources

Social media is collected in a number of ways, some of which make use of commercial search engines and targeted crawling, which have access potentially to the full Internet domain. For this reason an exhaustive list of “data sources” is not realistic.

However, a number of data sources of special importance were specifically targeted and specific parsers were written for them. These are summarized in the following Table 6:

Table 6: list of newspapers selected for data extraction in D2.2

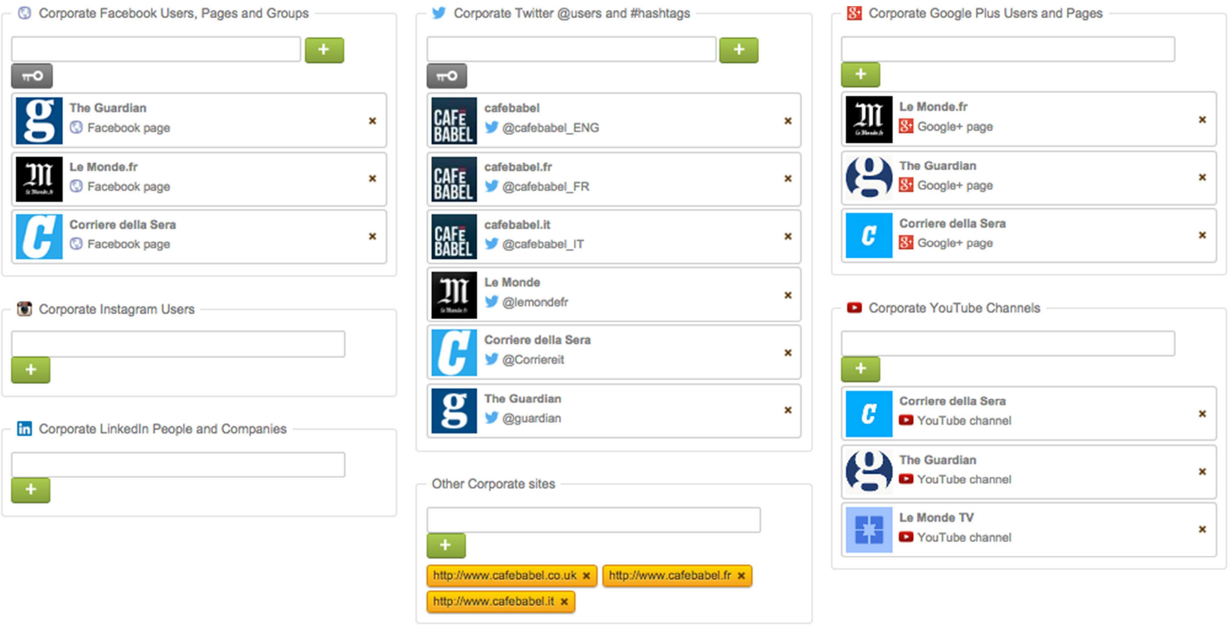


Country	Type	Name	URL
English	News	The Guardian	http://www.theguardian.com/uk
English	News	The Independent	http://www.independent.co.uk/
English	News	The Standar	http://www.standard.co.uk/
France	News	La Provence	http://www.laprovence.com/
France	News	Le Figaro	http://www.lefigaro.fr/
France	News	Le Monde	http://www.lemonde.fr/
France	News	L'Express	http://www.lexpress.fr/
France	News	Les Echos	http://www.lesechos.fr/
France	News	Libération	http://www.liberation.fr/
France	News	20 Minutes	http://www.20minutes.fr/
France	News	Metronews	http://www.metronews.fr
Italy	News	Corriere della Sera	http://www.corriere.it/
Italy	News	Metronews	http://www.metronews.it
Italy	News	Il Messaggero	http://www.ilmessaggero.it/
Spain	News	El Mundo	http://www.elmundo.es/
Spain	News	El País	http://elpais.com/
Spain	News	El Periodico	http://www.elperiodico.com/
Italy	News-blog	Republicca: vitorio zucconi	http://zucconi.blogautore.repubblica.it/

These are the same sources already targeted in D2.1, with some additions such as CaféBabel in English, French and Italian. For each of these sources, an example page with comments can be found at

http://5.9.95.170/dokuwiki/doku.php?id=sensei_data:data_requirements&#newspapers

Furthermore, we collected all content published by some social media channels (Twitter, Google+, Youtube and Facebook) of these newspapers, in particular those shown in the following Figure 5.



Corporate Facebook Users, Pages and Groups

- The Guardian – Facebook page
- Le Monde.fr – Facebook page
- Corriere della Sera – Facebook page

Corporate Instagram Users

Corporate LinkedIn People and Companies

Corporate Twitter @users and #hashtags

- cafebabel – @cafebabel_ENG
- cafebabel.fr – @cafebabel_FR
- cafebabel.it – @cafebabel_IT
- Le Monde – @lemondefr
- Corriere della Sera – @Corriereit
- The Guardian – @guardian

Other Corporate sites

- http://www.cafebabel.co.uk
- http://www.cafebabel.fr
- http://www.cafebabel.it

Corporate Google Plus Users and Pages

- Le Monde.fr – Google+ page
- The Guardian – Google+ page
- Corriere della Sera – Google+ page

Corporate YouTube Channels

- Corriere della Sera – YouTube channel
- The Guardian – YouTube channel
- Le Monde TV – YouTube channel

Figure 5: Social Media channels for which all published content was collected.

(Other social media content was collected by topic.)

3.4 Topics and Use Cases

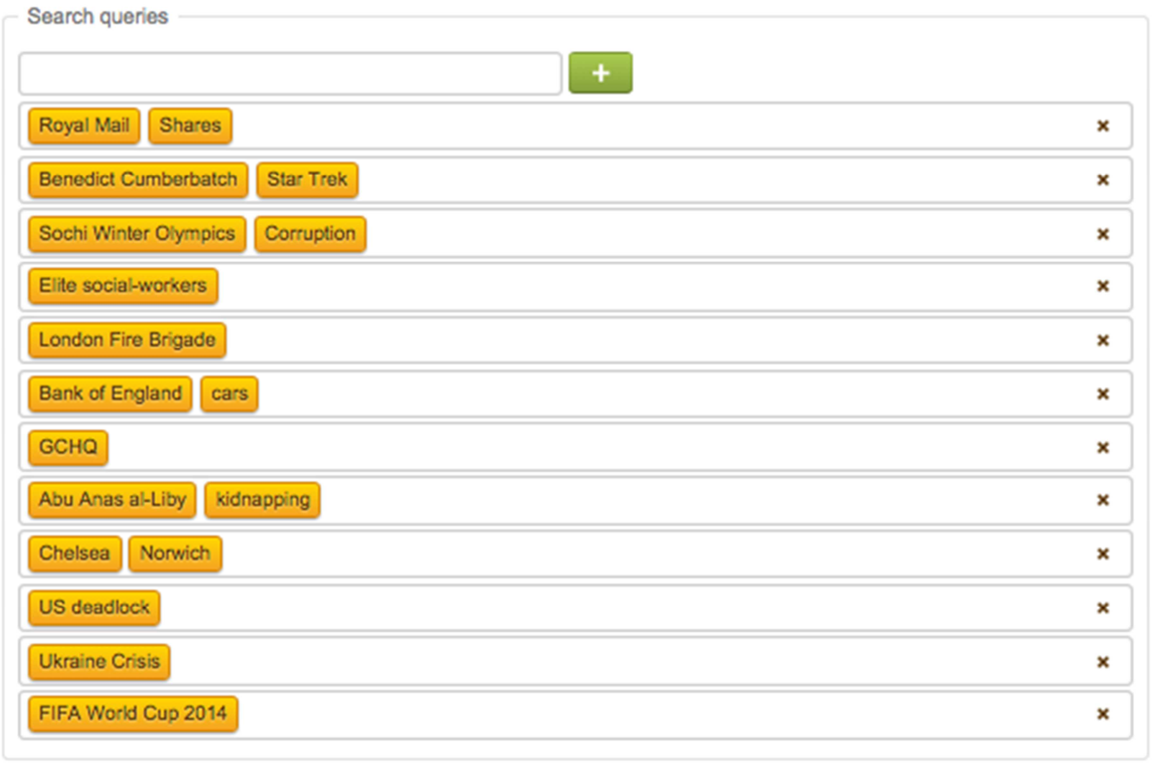
Different queries were used to collect data potentially relevant to the different social media use cases. Initial (D2.1) the topical query: “Europe OR Europa” was used to obtain documents from each newspaper source, resulting in over 3000 clippings (taking comments into account).

Besides these high-profile newspapers, in order to obtain mentions form general blogs and forums, we issued the query “Europe OR Europa” to Google Search using the news, forums and blogs filters for the time-periods of “this month” and “this year” (the queries were executed in March 2014), resulting in six paginated queries and over 5000 mentions. A total of 350k items where collected in this manner.

After D2.1 these queries were extended to collect data about other topics that were thought potentially relevant to the project. The Websays Dashboard graphical interface allows partners to add queries temporarily to collect data relevant to a topic or subtopic interesting for investigation. Throughout the development of the use cases, partners have been able to tune queries and browse results. All data collected is added to the data collection, which is dumped periodically for all partners to access. The main topics of enquiry were:

- RATP (Paris public transportation system) conversation about the brand in french.
- Orange (Telephonie company) conversation about the brand in french.
- FIFA World Cup 2014: conversation about the world cup world-wide
- Other hot topics: GCHQ, Abu Anas al-Liby, Ukraine Crisis, Sochi Winter Olympics, etc.

For illustration, we present the queries currently configured for crawling under one of SENSEI's crawl profiles, where we can see different topics of current exploration:



Search queries

+

Royal Mail | Shares ×

Benedict Cumberbatch | Star Trek ×

Sochi Winter Olympics | Corruption ×

Elite social-workers ×

London Fire Brigade ×

Bank of England | cars ×

GCHQ ×

Abu Anas al-Liby | kidnapping ×

Chelsea | Norwich ×

US deadlock ×

Ukraine Crisis ×

FIFA World Cup 2014 ×

Figure 6: An illustration of a set of queries activated for targeted crawling using the Websays Dashboard graphical interface.

In Figure 6 each row corresponds to a separate query (combined disjunctively) and within each row, boxes represent phrases and are combined conjunctively.

3.5 Content Extraction

Content extraction is composed of these three steps. Each requires customization to tackle specifically formatted data sources, and required the development of modules for each of the sources listed above:

- 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.

We wrote specialized parsers for each newspaper to convert the newspaper page contents into the Data Schema reported above. In some cases, when newspapers use HTML IFrames to display comments, it was necessary to build distinct parsers for the post and the comments.

3.6 Pre-processing

The Websays pre-processing pipeline was applied to all the documents entering the SENSEI collection index (via targeted crawling, specific document demands via the asynchronous crawler, external APIs, etc.). The pre-processing of WEB data is also described in D3.1, because it's also part of activity 3.1.

We highlight here the main components of pre-processing. :

- **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.
- **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.

3.7 Evaluation

Evaluation of the data collection process was carried out by analysts and developers at Websays and at UNITN, USFD and AMU in order to detect:

- Missing content (such as missing URLs or articles, missing comments within the articles, etc.)
- Missing metadata (such as missing authors, dates, scores, comments, comment references, etc.)
- Systematic pre-processing errors. Due to the nature of pre-processing it is acceptable to have many labeling errors. However, systematic errors and biases were sought and corrected.

All manual errors were marked using the Websays Dashboard annotation tool and corrected. Furthermore, the entire collection was reprocessed multiple times when systematic errors discovered led to improvements in the pre-processing pipeline.

3.8 Data characteristics

The D2.2 collection contains over 4 million posts, and over 1.5 million conversations with more than one post. Posts come from thousands of different domains, including blogs, forums, and multiple social media channels, and are written in hundreds of languages (although most posts are in English, French and Italian).

The main topics of the posts in the collection are (as discussed in section 3.4): News Hot Topics, Specific NewsPaper Social Media Publications, RATP and Orange.

As an example of the characteristics of sub-collections, we give statistics of the first two sub-collections.

General News and Newspaper Social Media Publications:

- Size: 4.4M posts, 1.1M “parent” posts (not counting comments, retweets, etc.) and 3.3M “comment” posts
- Most frequent domains (and number of posts per domain):

www.twitter.com:2815993
www.facebook.com:1351671

www.youtube.com:53543
plus.google.com:41093
discussion.theguardian.com:26654
www.theguardian.com:7028
www.lemonde.fr:4009
www.independent.co.uk:3343
www.corriere.it:3171
www.reuters.com:1169
www.newslocker.com:1142
www.pinterest.com:1062
timesofindia.indiatimes.com:992
vimeo.com:773
instagram.com:610
bootstrap.liberation.fyre.co:570
www.reddit.com:457
time.com:446
www.bbc.co.uk:426
sports.ndtv.com:409

- Most frequent author location strings:

London:87780
Paris:63306
France:34637
Milano:28755
Uk:25991
Seattle, Wa:17317
Mexico:14807
Rawalpindi:12206
Liverpool, United Kingdom:11473
Roma:10792
London, Uk:10737
Worldwide:8732
Italia:8704
India:8292
Paris, France:8257
Usa:7947
United Kingdom:7863
Reino Unido:7739
Italy:7091
England:6990

- Most frequent languages detected:

English:2429469
Italian:830712
French:815567
UNKNOWN:143638
Spanish:32850
German:25174
Malay:12990

Indonesian:10910
Portuguese:10504
Russian:9847

RATP:

- Size: 118k posts, 59k “parent” posts (not counting comments, retweets, etc.) and 58k “comment” posts.
- Most frequent domains (and number of posts per domain):

  www.twitter.com:94659
  www.facebook.com:13204
  instagram.com:2535
  www.youtube.com:1456
  www.ratp.fr:322
  www.lefigaro.fr:216
  www.lemonde.fr:190
  www.wizbii.com:165
  www.liberation.fr:127
  www.leparisien.fr:114
  vimeo.com:108
  www.vianavigo.com:104
  plus.google.com:101
  fr.news.yahoo.com:60
  www.20minutes.fr:54
  ask.fm:52
  premiersmetros.tumblr.com:43
  www.blogencommun.fr:36
  www.rtl.fr:35
  tempsreel.nouvelobs.com:34

- Most frequent author location strings:

  Paris:15380
  France:5367
  Mexico:757
  Lyon:349
  Francia:253
  Ile De, France:250
  Bordeaux:248
  Lille:236
  Marseille:198
  Toulouse:195
  Paris France:188
  Nantes:186
  France, Paris:151
  France, Idf:140
  Paris Provins Meaux, Coulom:134
  Paris, Ile De France:127
  Strasbourg:123
  Paris, Los Angeles:117

Ile De France:116

- Most frequent languages detected:

  French:99914
  English:7236
  UNKNOWN:6127
  Spanish:1408
  Italian:638
  Romanian:498
  German:184
  Portuguese:143
  Arabic:131
  Dutch:116

In Appendix B is presented the current state of data collection and annotation.

3.9 Crawler Adaptation

Websays provided its high performance crawler, processing and indexing platform. However a number of adaptations were necessary: , the main ones being:

- Online (non-batch, on demand) crawler: an asynchronous fetcher and processor was developed to allow the crawling of any URL in online mode. This is discussed below in more detail. This crawler can now be accessed by two mechanisms:
  - REST requests
  - WebApp
- Query-agnostic crawling and parsing: Previously Websays only parsed segments of documents matching specific queries. In order to parse and index entire threads and page collections, a query-agnostic segmenter and parser were developed.

In order to index on-demand specific URLs under investigation, Websays developed an asynchronous crawler with a fast queue so that partners could at any time request URLs to be fetched, indexed and added to the SENSEI collection.

The following Figure gives an overview of the architecture behind this service, which was already presented in D2.1.

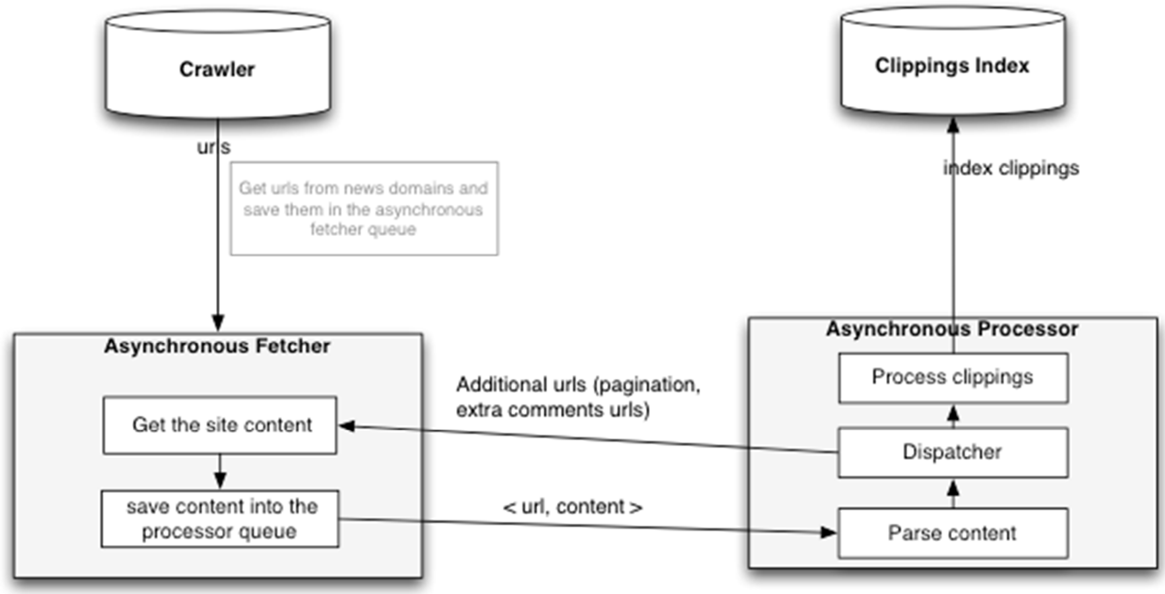
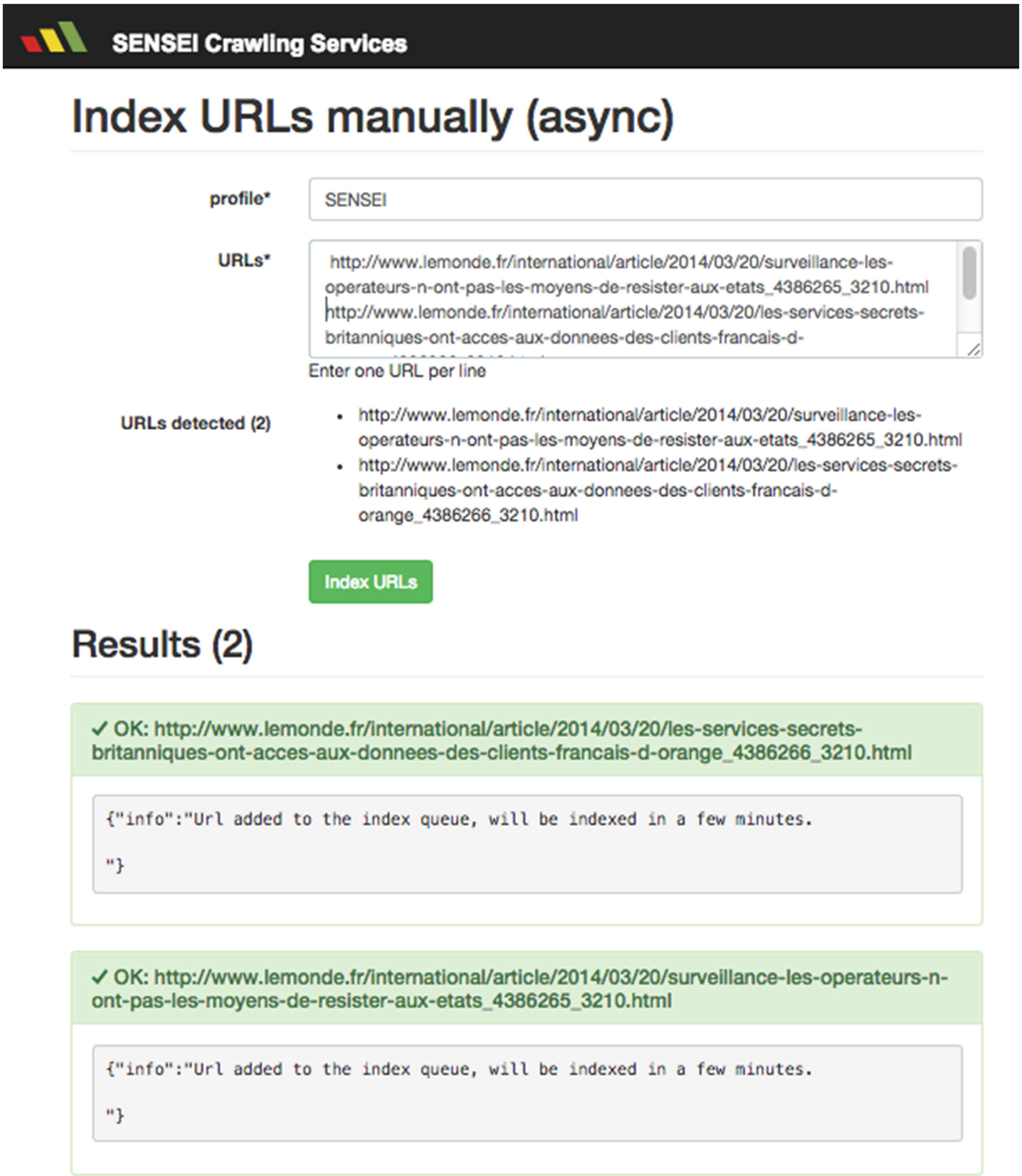

Crawler
urls
Get urls from news domains and save them in the asynchronous fetcher queue
Asynchronous Fetcher
Get the site content
save content into the processor queue
Additional urls (pagination, extra comments urls)
< url, content >
Clippings Index
index clippings
Asynchronous Processor
Process clippings
Dispatcher
Parse content

Figure 7: Asynchronous Fetcher developed to serve online fetch requests within an agile parser development environment.

This crawler has a REST interface so that partners can invoke it programmatically. A visual interface has also been developed (a web app) to allow partners to request URLs manually, without having to program. An example execution with this WebApp is shown in the following Figure 8.



Index URLs manually (async)

profile* SENSEI

URLs* http://www.lemonde.fr/international/article/2014/03/20/surveillance-les-operateurs-n-ont-pas-les-moyens-de-resister-aux-etats_4386265_3210.html
http://www.lemonde.fr/international/article/2014/03/20/les-services-secrets-britanniques-ont-acces-aux-donnees-des-clients-francais-d-

Enter one URL per line

URLs detected (2)

- http://www.lemonde.fr/international/article/2014/03/20/surveillance-les-operateurs-n-ont-pas-les-moyens-de-resister-aux-etats_4386265_3210.html
- http://www.lemonde.fr/international/article/2014/03/20/les-services-secrets-britanniques-ont-acces-aux-donnees-des-clients-francais-d-orange_4386266_3210.html

Results (2)

✓ OK: http://www.lemonde.fr/international/article/2014/03/20/les-services-secrets-britanniques-ont-acces-aux-donnees-des-clients-francais-d-orange_4386266_3210.html

```
{"info":"Url added to the index queue, will be indexed in a few minutes.

"}
```

✓ OK: http://www.lemonde.fr/international/article/2014/03/20/surveillance-les-operateurs-n-ont-pas-les-moyens-de-resister-aux-etats_4386265_3210.html

```
{"info":"Url added to the index queue, will be indexed in a few minutes.

"}
```

Figure 8 Usage example of the SENSEI crawling on-demand service.

3.10 Social Media Annotation

For use in the intrinsic evaluation of social media summaries (see D1.2) and in order to fulfill Milestone 1 “Annotated data sets for evaluation”, the USFD team has collected news articles and their associated comments from *The Guardian* and has annotated them with human authored summaries.

Data

For this annotation task we used 20 articles. We took 5 articles from each of the domains: “World News”, “UK News”, “Environment” and “Business”. The articles were randomly selected. Comments on news articles in *The Guardian* are organized into threads – a starting comment and sub-comments below the starting comment. When selecting the articles for the annotation task we stipulated that to be selected an article had to have at least 100 comments (sum of starting and sub-comments). For each article we also collected the (chronologically) first *k* threads such that at least 100 comments were gathered (the number of threads needed to meet this condition varies per article). This means that for some articles more than 100 comments were collected. On average the 20 articles gathered each have 105 associated comments. The maximum number of comments is 127. In total, the dataset contains 119,689 words – words counted from the article and the comments (i.e. excluding annotations) and comprises 265 “dialogues” (here we count each thread as a dialogue, as specified in the Milestone).

Tool support

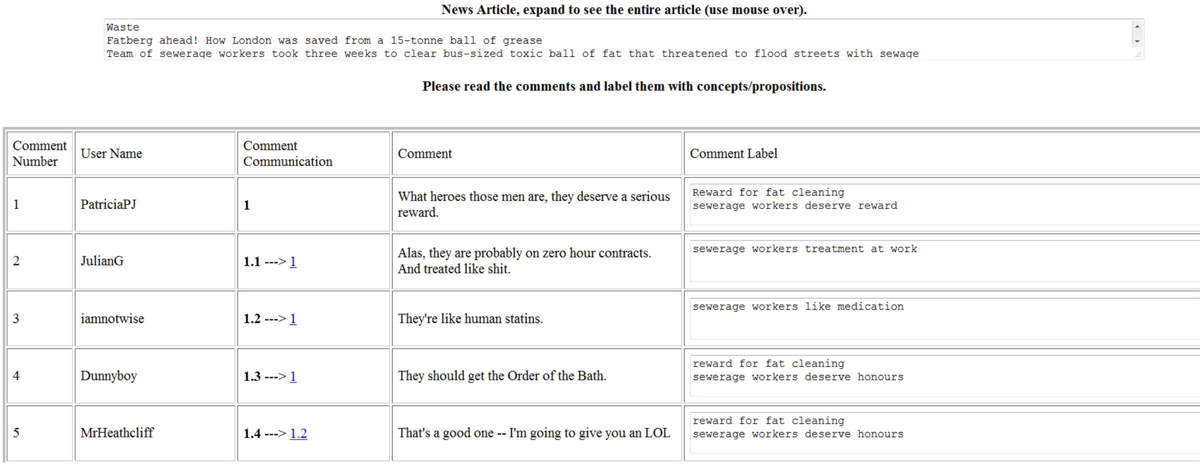
To support annotators in the summary writing task (described in D1.2, Section 4.2.2) we have developed an interface that i) allows an annotator to select an article and comment set for annotation and then ii) displays the article and the associated set of reader comments, preserving the thread structure and original user ids. For each comment the interface provides a cell to hold a label annotation for that comment (Figure 9). There is also a text box for collecting labels.

In addition we advised annotators to use a text editor of their choice for gathering labels and comments, for reformulating labels, for quantifying ideas, for note taking, saving interesting comments and labels, and for generating the written summary.

The final summary is returned via a text box in the interface (Figure 10). Annotators may also supply an “unconstrained summary”, which is an initial “natural-length” summary they may produce without worrying about the 150-250 length word constraint.

Annotation Process

The annotation was performed by five people who were members of the USFD SENSEI team. To prepare annotators for the summary writing task, we presented an overview of our guidelines for writing summaries of reader comments, with examples; the presentation was followed by a training session in which we asked annotators to carry out various exercises based on the method; these were designed to help annotators practice writing summaries and to practice using the supporting interface. In a final exercise, we asked annotators to produce a summary for a sample article and comment set.



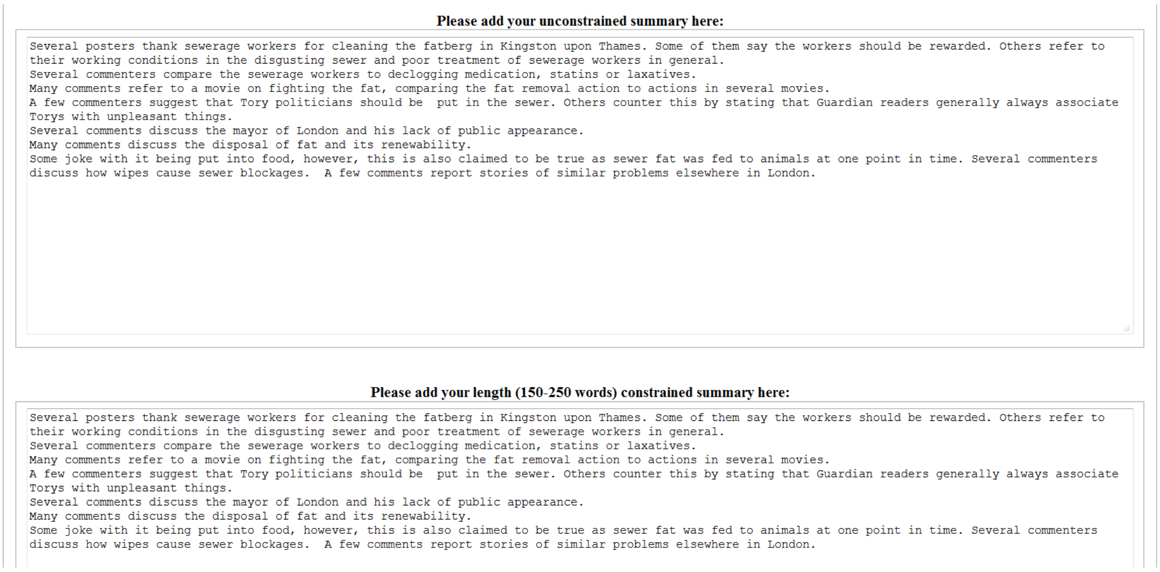
News Article, expand to see the entire article (use mouse over).

Waste
Fatberg ahead! How London was saved from a 15-tonne ball of grease
Team of sewerage workers took three weeks to clear bus-sized toxic ball of fat that threatened to flood streets with sewage

Please read the comments and label them with concepts/propositions.

Comment Number	User Name	Comment Communication	Comment	Comment Label
1	PatriciaPJ	**1**	What heroes those men are, they deserve a serious reward.	Reward for fat cleaning sewerage workers deserve reward
2	JulianG	**1.1** ---> 1	Alas, they are probably on zero hour contracts. And treated like shit.	sewerage workers treatment at work
3	iamnotwise	**1.2** ---> 1	They're like human statins.	sewerage workers like medication
4	Dunnyboy	**1.3** ---> 1	They should get the Order of the Bath.	reward for fat cleaning sewerage workers deserve honours
5	MrHeathcliff	**1.4** ---> 1.2	That's a good one -- I'm going to give you an LOL	reward for fat cleaning sewerage workers deserve honours

Figure 9: Comments and labels



Please add your unconstrained summary here:

Several posters thank sewerage workers for cleaning the fatberg in Kingston upon Thames. Some of them say the workers should be rewarded. Others refer to their working conditions in the disgusting sewer and poor treatment of sewerage workers in general.
Several commenters compare the sewerage workers to declogging medication, statins or laxatives.
Many comments refer to a movie on fighting the fat, comparing the fat removal action to actions in several movies.
A few commenters suggest that Tory politicians should be put in the sewer. Others counter this by stating that Guardian readers generally always associate Torys with unpleasant things.
Several comments discuss the mayor of London and his lack of public appearance.
Many comments discuss the disposal of fat and its renewability.
Some joke with it being put into food, however, this is also claimed to be true as sewer fat was fed to animals at one point in time. Several commenters discuss how wipes cause sewer blockages. A few comments report stories of similar problems elsewhere in London.

Please add your length (150-250 words) constrained summary here:

Several posters thank sewerage workers for cleaning the fatberg in Kingston upon Thames. Some of them say the workers should be rewarded. Others refer to their working conditions in the disgusting sewer and poor treatment of sewerage workers in general.
Several commenters compare the sewerage workers to declogging medication, statins or laxatives.
Many comments refer to a movie on fighting the fat, comparing the fat removal action to actions in several movies.
A few commenters suggest that Tory politicians should be put in the sewer. Others counter this by stating that Guardian readers generally always associate Torys with unpleasant things.
Several comments discuss the mayor of London and his lack of public appearance.
Many comments discuss the disposal of fat and its renewability.
Some joke with it being put into food, however, this is also claimed to be true as sewer fat was fed to animals at one point in time. Several commenters discuss how wipes cause sewer blockages. A few comments report stories of similar problems elsewhere in London.

Figure 10: Constrained and Unconstrained Summaries.

Examples

Here is an example reader comment and labels supplied by annotators:

Comment:

> Have you seen BJ since that interview? Nope, me neither. I've a feeling he's just been jet-washed off his favourite hiding place.

Labels identified:

> London mayor interview;
>
> London mayor doesn't appear in public after interview on fat reuse;

Figure 11 contains an example summary of a comment set created according to our method.

Several posters thank sewerage workers for cleaning the fatberg in Kingston upon Thames. Some of them say the workers should be rewarded. Others refer to their working conditions in the disgusting sewer and poor treatment of sewerage workers in general.

Several commenters compare the sewerage workers to declogging medication, statins or laxatives.

Many comments refer to a movie on fighting the fat, comparing the fat removal action to actions in several movies.

A few commenters suggest that Tory politicians should be put in the sewer. Others counter this by stating that Guardian readers generally always associate Tories with unpleasant things.

Several comments discuss the mayor of London and his lack of public appearance.

Many comments discuss the disposal of fat and its renewability.

Some joke with it being put into food, however, this is also claimed to be true as sewer fat was fed to animals at one point in time. Several commenters discuss how wipes cause sewer blockages. A few comments report stories of similar problems elsewhere in London.

Figure 11: Example News Comment Summary

4. Conclusion

In the D2.2 deliverable, a subset of the original speech corpus has been selected and annotated. During the definition of Use Cases and Evaluation methodology, it has been necessary to develop new tools and customize existing ones, in order to collect the data required by other WPs.

For deliverable D2.3 there are several important aspects of data collection that need to be tackled. First, as use cases are consolidated and prototypes development and annotation begin, we will find to what degree the current data (both speech and social media) is sufficient for the project goals, and to what degree it needs to be complemented with new targeted crawls, new analysis and annotations, etc. Second, as prototypes take shape we will be able to evaluate if current pre-processing is sufficient or needs to be improved, etc. Finally, another important aspect for D2.3 is to achieve a greater degree of integration and inter-play between social media data and speech data, both at the level of representation and annotation as in terms of topical association. Again this will be naturally driven by the prototype work.

Appendix A: Social Media Data Schema

Data Schema

Data Type	Data Field Description
IDs	Integer **externalID** (unique ID of an object in the corpus) Integer **parentID** (externalID of parent object, e.g. the ID of the article if the object is a comment) Integer **postID** (externalID of the “parent” of the tree, e.g. the post originating the conversation) Integer **versionID** (ID for the version of edited posts) String **domain** (domain in the url of the post) String **APIObjectID**; (id of the post on the original media platform) String **APIAuthorID**; (the author ID in the native API where this comes from: e.g. Facebook user id) String **APIToAuthorID**; (ID in the native API of the author targeted by the message)
Type	String **postType** (article, comment, status update, reply, repost, etc) String **pageSuperType** (social, news, blog, forum, video, other) String **sourceType** (e.g. guardian, corriere, metronews.fr, facebook, twitter, etc.)
Post Data	String **title**; (title of the post – if available) String **keywords**; (keywords of the post – if available) String **text**; (body of the post) String **textHTML**; (raw page) String **author**; (author of the post, surface username)
Metadata	Integer **page_numOfComments**; (number of comments) Integer **page_numOfLikes**; (Likes in Facebook, News and blogs) Integer **page_numOfDislikes**; (Dislikes of youtube or metronews) Integer **page_numOfViews**; (Youtube views) Integer **page_numOfFavorites**; Twitter Favorites) Integer **page_numOfReTweets**; (ReTweets) Integer **page_numOfShares**; (Shares in facebook or g+) String **mood** (mood type in corriere.it) Integer **moodStrength** (mood strength in corriere.it) String **inReplyTo** (adressee of comments) String **embeddedMediaType** (text, text+photo, text+video, text+link, photo, video, link) List<String> **authorSource**; (from re-tweets, etc.) String **tags**; (tags of the post – if available) Boolean **isBestComment** (label for “guardian picks” and best comments) String **pictureURLs** (URL of pictures in articles or posts) String **mediaURLs** (URL of video or other media included in articles or posts)
Author Metadata	String **authorType**; (anonymus, user, group, UKNOWN) String **authorProfilePictureURL** (url of the user profile picture) Integer **user_numOfFollowers**; (Followes of twitter) Integer **user_numOfFollowing**; (Following in Twitter) Integer **user_numOfFriends**; (Friends of facebook and other social) List<String> **authorsMentioned**; (other authors mentioned in the post)
Preproc. Annotations	LanguageTag **langDetected** (language of match automatically detected); LanguageTag **langReported** (language of the social interface displayed to the user); Integer **websaysPolarity** (sentiment analysis); List<String> **clusters** (clustering together similar posts based on matchKeywords, matchAuthorNames or Polarity);

Time-stamps	Date **date**;(clipping date real or extracted or guessed) Date **indexingTimeStamp**; (indexing date) Integer **timeRank** (date/time-based ranking for visualization e.g. 1,2,3,4 from the oldest to the newest)
Matches	String **crawlQueryMatch** (text that matched the query that triggered crawling this clipping) Integer **crawlQueryID** (id of queries that triggered crawling this clipping)
Localiza-tion	String **authorLocation**; (geo-location of the author)
Metrics	Integer **count1** (overall impact (e.g. number of re-tweets));
NLP	String **nlp_chunk** (shallow chunks annotation); String **nlp_pos** (shalow part of speech annotation);

Resulting XML Schema

```
<xs:schema attributeFormDefault="unqualified" elementFormDefault="qualified"
targetNamespace="http://websays.com/"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="senseiClipping">
 <xs:complexType>
 <xs:sequence>
 <xs:element type="xs:int" name="externalID"/>
 <xs:element type="xs:int" name="postID"/>
 <xs:element type="xs:int" name="parentID"/>
 <xs:element type="xs:int" name="versionID"/>
 <xs:element type="xs:string" name="sourceID"/>
 <xs:element type="xs:string" name="authorID"/>
 <xs:element type="xs:string" name="domain"/>
 <xs:element type="xs:string" name="apiObjectID"/>
 <xs:element type="xs:string" name="apiAuthorID"/>
 <xs:element type="xs:string" name="apiToAuthorID"/>
 <xs:element type="xs:string" name="superType"/>
 <xs:element type="xs:string" name="type"/>
 <xs:element type="xs:string" name="sourceType"/>
 <xs:element type="xs:string" name="authorType"/>
 <xs:element type="xs:string" name="tittle"/>
 <xs:element type="xs:string" name="keywords" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:string" name="text"/>
 <xs:element type="xs:string" name="textHTML"/>
 <xs:element type="xs:int" name="numOfComments"/>
 <xs:element type="xs:int" name="numOfLikes"/>
 <xs:element type="xs:int" name="numOfDislikes"/>
 <xs:element type="xs:int" name="numOfViews"/>
 <xs:element type="xs:int" name="numOfFavorites"/>
 <xs:element type="xs:int" name="numOfReTweets"/>
 <xs:element type="xs:int" name="numOfShares"/>
 <xs:element type="xs:string" name="author"/>
 <xs:element type="xs:string" name="authorProfilePictureURL"/>
 <xs:element type="xs:int" name="user_numOfFollowers"/>
 <xs:element type="xs:int" name="user_numOfFollowing"/>
 <xs:element type="xs:int" name="user_numOfFriends"/>
 <xs:element type="xs:string" name="authorsMentioned" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:string" name="authorSource" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:string" name="mood"/>
 <xs:element type="xs:int" name="moodStrength"/>
 <xs:element type="xs:string" name="inReplyTo"/>
 <xs:element type="xs:string" name="embeddedMediaType"/>
 <xs:element type="xs:string" name="tags"/>
 <xs:element type="xs:string" name="isBestComment"/>
 <xs:element type="xs:string" name="pictureURLs" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:string" name="mediaURLs" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:dateTime" name="date"/>
 <xs:element type="xs:dateTime" name="indexingTimeStamp"/>
 <xs:element type="xs:int" name="timeRank"/>
 <xs:element type="xs:string" name="crawlQueryMatch"/>
 <xs:element type="xs:int" name="crawlQueryID"/>
 <xs:element type="xs:string" name="authorLocation"/>
 <xs:element type="xs:string" name="langDetected"/>
 <xs:element type="xs:string" name="langReported"/>
 <xs:element type="xs:int" name="websaysPolarity"/>
 <xs:element type="xs:string" name="clusters" maxOccurs="unbounded" minOccurs="0"/>
 <xs:element type="xs:int" name="count1"/>
 <xs:element type="xs:int" name="nlp_chunk"/>
 <xs:element type="xs:int" name="nlp_pos"/>
 </xs:sequence>
 </xs:complexType>
 </xs:element>
</xs:schema>
```

Appendix B: State of Data Collections

Size of Speech Data per Language

Speech Data Sets Available	TP Annotation Activity (M1-12)	AMU Annotation Activity (M1-12)	UNITN Annotation Activity (M1-12)	Language
572 LUNA dialogs	200 different dialogs have been annotated with AOF included segment turn and COF.	--	200 selected dialogs have been annotated with AOF and synopsis	Italian
1500 DECODA Conversations	118 different conversations have been annotated with AOF included segment turn and COF.	200 conversations have been annotated with at least two synopses	--	French

Size of Social Data Sets Per Language

Type	Language	N. of Dialogues or posts	N. of tokens
Social Media, News and Blogs	English	2.4M	>240M
Social Media, News and Blogs	French	.8M	>80M
Social Media, News and Blogs	Italian	.8M	>80M
Social Media, News and Blogs	Spanish	.3M	>30M
Social Media, News and Blogs	Other Languages	.2M	>20M

Appendix C: SENSEI ACOF Data Model

The data model of the application is composed of seven tables, as illustrated in figure 9 below.

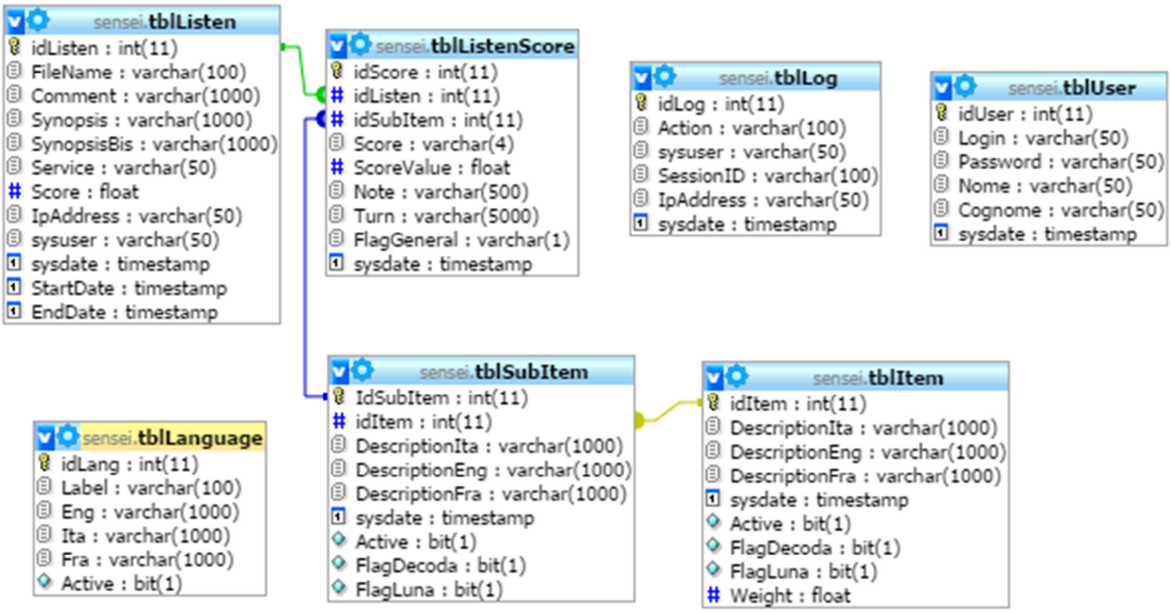


Figure 11: Data model of SENSEI Web Annotation tool

Table 7 below describes the structure of tblUser, which contains the users enabled to access to the application.

Table 7: Table tblUser

Field	Description
idUser	Unique identifier of the user
Login	Username of user
Password	Password of user
Nome	Firstname of user
Cognome	Lastname of user
sysdate	Date,hour and minutes of system

Table 8 below describes the structure of tblLanguage, which contains the translation of the labels of the user interface in the different languages.

Table 8: tblLanguage

Field	Description
idLang	Unique identifier of the Language
Label	Name of the label
Eng	Italian translation of this label
Ita	English translation of this label
Fra	French translation of this label
Active	Status of label

Table 9 below describes the structure of tblListen, which contains the main information of the AOF and COF.

Table 9: tblListen

Field	Description
idListen	Unique identifier of the monitor
FileName	Transcription Filename
Comment	Text field where the quality assurance professional write generic comments
Synopsis	Text field where the quality assurance professional write Synopsis comments
SynopsisBis	Text field where the quality assurance professional write other Synopsis comments
Service	Type of service(LUNA,DECODA)
Score	Score value of monitor
IpAddress	Address of user machine
sysuser	Current username
sysdate	Date, hour and minutes of system

Table 10 below describes the structure of tblListenScore, which contains the score of each subitem of the AOF and the relevant segment-turn.

Table 10: tblListenScore

Field	Description
idScore	Unique identifier of the score
idListen	Unique identifier of the monitor
idSubItem	Unique identifier of the subItem
Score	Score of subitem(PASS,FAIL,NA)
ScoreValue	Score value of subitem
Note	Text field where the quality assurance professional write subitem notes
Turn	Text field with the segment turn selected, the start startTime and the endTime taken from the tag <Turn > of transcription file
FlagGeneral	Flag General equal Y indicate that there is no relevant speech turn for the item
sysdate	Date, hour and minutes of system

The table 11 below describes the structure of tblItem, which contains the item of the AOF.

Table 11: tblItem

Field	Description
idItem	Unique identifier of the Item
DescriptionIta	Italian description of this item
DescriptionEng	English description of this item
DescriptionFra	French description of this item
sysdate	Date, hour and minutes of system
Active	Status of item
FlagDecoda	Flag Service
FlagLuna	Flag Service
Weight	Weight of the item

Table 12 below describes the structure of tblsubitem, which contains the sub-item of the AOF.

Table 12: subitem

Field	Description
IdSubItem	Unique identifier of the SubItem
idItem	Unique identifier of the Item
DescriptionIta	Italian description of this SubItem
DescriptionEng	English description of this SubItem
DescriptionFra	French description of this SubItem
sysdate	Date, hour and minutes of system
Active	Status of SubItem
FlagDecoda	Flag Service
FlagLuna	Flag Service