HAL open science

Delivrable D5.3 - Evaluation Report

Amel Fraisse, Patrick Paroubek, Xavier Tannier

To cite this version:

Amel Fraisse, Patrick Paroubek, Xavier Tannier. Delivrable D5.3 - Evaluation Report. 2014. hal-01933447

HAL Id: hal-01933447

https://hal.science/hal-01933447

Submitted on 22 Dec 2020

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

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

Deliverable D5.3 - Evaluation Report

LIMSI-CNRS

WP5 (T5.3)

Historique des modifications			
Version	**Auteur**	**Date**	**Description des modifications**
1	Xavier Tannier Patrick Paroubek	May. 22th 2014	Annotation representation Model
2	Amel Fraisse	Jul. 20th 2014	Lexical Ressources Evaluation
3	Patrick Paroubek	Jul. 22th 2014	Evaluation Infrastructure
4	Amel Fraisse	Oct. 06th 2014	Expert Annotation Guidelines
5	Patrick Paroubek	Nov. 06th 2014	State of the Art on Crowdsourcing

Validation			
Role	**Organisation**	**Name**	**Date**

Table of Content

1 Introduction **4**
1.1 Aim 4
1.2 Responsibility 4

2 Affect Lexica **4**
2.1 Aim 4
2.2 State of the art 4
2.2.1 Semantics 5
2.2.2 Opinion Mining & Sentiment Analysis 9
2.2.3 Named Entities 10
2.2.4 Machine Translation 11
2.2.5 Information Retrieval/Extraction & Question Answering 13
2.2.6 Language Communication Enhancement/Validation 14
2.2.7 Real Time Dialog System 15
2.2.8 Conclusion 16
2.3 7 language lexica Experiment 16
2.3.1 The HC experiment with CrowdFlower 17

3 Evaluation Campaign **21**
3.1 Aim 21
3.2 Data 21
3.3 A model of language data annotation 22
3.3.1 Annotation representation 22
3.4 Applying our model to real evaluations 24
3.4.1 Classification 25
3.4.2 Transduction 29
3.5 Expert Annotation Guidelines 30
3.5.1 Groups 31
3.5.2 Relations 42
3.6 Expert Annotation Software 45
3.7 Campaign Deployment 45

Index of Figures
Index of tables

1 Introduction

1.1 Aim

This document describes the design and execution of experiments for assessing HC effectiveness. It constitutes the first version of the deliverable planned for the end of the project, as such it focuses on experiments design and present only results for what has been achieved so far (T5.1 and part of T5.2). The whole set of evaluation results, including the results of the evaluation campaign (T 5.3) planned to happen during the last year of uComp will be presented in the final devliverable, due for M30.

1.2 Responsibility

The partner responsible the production of this document is LIMSI-CNRS.

2 Affect Lexica

2.1 Aim

The task for *HC-Based Knowledge Resource Evaluation* focuses on the evaluation of the HC-sourced affect lexicons both in terms of improving the recall, precision and F-measure of existing sentiment and opinion mining algorithms ([Gindl et al., 2010] and [Pak et al., 2014]), as well as by using a set of additional games aimed at evaluating specific aspects of these resources.

We plan to assess the interest of replacing/complementing entropy-based data browsing algorithms by/with human guidance. This will help evaluate the effectiveness of the HC paradigm and framework, and address the following questions:

- *How do crowdsourced resources compare to those created manually, automatically or through mechanised labour (in terms of both quantity and quality, which will be measured by the k-value of inter-annotator agreement)?*
- *Is the cost/benefit ratio of this process better than that of other knowledge acquisition approaches?*

Analogous to WP4, T5.2 will primarily focus on the climate change domain and the repository built in T1.5.

2.2 State of the art

Since it was created in 2005 according to [Safire, 2009], the word *crowdsourcing* found rapidly[1] its way into the research community, where people tried from the beginning to assess the benefits and gains they could get from using the power of human computation. Most of the early

[1] in 2009 it reached the 1 million hit limit in Google ([Safire, 2009])

studies reported in the litterature mention Amazon Mechanical Turk (AMT), with CrowdFlower appearing a little later. Our survey of previous work is organized along the following broad themes:

- Semantics,
- Opinion Mining & Sentiment Analysis,
- Named Entities,
- Machine Translation,
- Information Retrieval/Extraction & Question Answering,
- Language Communication Enhancement/Validation,
- and Real Time Dialog Systems.

2.2.1 Semantics

In [Rumshisky et al., 2009], the authors report on an experiment to build a fuzzy sense inventory database for a set of polysemous verbs of medium difficulty, using non-expert annotators which were presented with sentences containing one of the verb form and had identify particular meaning. The results were then evaluated against the groupings created by a professional lexicographer (set-matching F-score of 0.93), each sentence being annotated by 5 AMT annotators. On average 1 minute was required to process 10 sentences and the amount paid to the annotator was 0.03 USD. The total sum spent during the first experiment was 10$. Later, [Rumshisky, 2011] conclude that clustering the 350 concordance lines into sense related groups would yield only reliable results for about 140 concordance lines and [Rumshisky et al., 2012] report that one of the major issue is to attract high-quality annotators on a service like AMT to perform complex linguistic tasks. If the use of best practice guidelines helps, it does not supresses the need to run preliminary experiments to calibrate the task parameters and interface. Another quality improvement in the result can be obtained by comparing AMT worker results using worker-quality weighted majority votes. However, while writing about the 2009 experiement [Rumshisky et al., 2012] report that "*the experiments run with the same parameters today do not lead to either fast completion or quality annotation.*". The authors found found they could get an important increase of results quality by restricting the location of the annotators to the USA, but the tak took longer to complete and costed more. Factors that attract good annotators in large numbers are the pay rate (which can varied at that time from a few cents to 20 $ per task), the height of the task in the task search space proposed to annotators and the apparent simplicity of the task and the clarity of the task description, as well as the standing of the image of the task proposer in the AMT annonators specific social media which depends on the rapidity of the payment and speed of answering AMT annotators' questions. For instance [Rumshisky et al., 2012] report that in the US local restricted prototype experiments, the task failed to complete when only 0.01 $ were offered to provide five judgments and the location of the taks was very low in the task list because not much data was proposed for

annotation. Keeping the in US requirement, the best strategy that the authors found was to pay 3 cents for 10 judgments in one task, allowing a maximum worker error rate of 27%, with at least a 85% previous task approval rate on at least 200 tasks, and offering bonuses to the top workers. They obtained an F-score of 0.72 with a kappa of 0.69. The differences observed by the authors between the 2009 and 2012 experiments show that the AMT market place is evolving rapidly and that the validity of best practice guidelines needs to be checked frequently.

In [Munro et al., 2010], the authors investigate the use of AMT for a range of linguistic experiments from semantics to psycholinguistics dealing with:

1. verbs semantics,
2. segmentation of audio speech stream,
3. language models,
4. speech grammaticality,
5. thematic role,
6. methapor brain processing,
7. and reading attentiveness.

For instance they report between crowdsourced and laboratory condition kappa values of 0.9 for the verb semantics task and of 0.759 for the language model task. Following their experimental results, they conclude that crowdsourcing provides the means to run systematic, large-scale judgment studies at a lower cost and much more easily than when doing them under laboratory conditions.

The experiment with AMT presented in [Negri et al., 2011] addresses sentence modification and textual entailment annotation in a multiligual context involving English, Italian and German. From a set of aligned sentences in the three languages, modifications are done in a monolignual set-up (English) by paraphrasing and rephrasing with information addition or deletion. Modified sentences are then translated after having been annotated with entailment information with respect to the source sentence. The process of multilingual entailment is then simplified by having entailment processing done in a monoligual set-up. For quality control, only the annotation that gathered the agreement of 4 out of 5 annotators were retained and gold standard data test were run. Six types of tasks were submitted to the AMT annotators:

1. paraphrasing,
2. grammaticallity checking,
3. bidirectionnal entailment,

4. rewriting while adding information,
5. rewriting while removing information,
6. unidirectional entailment,

for a total of 1,620 pairs of sentences added to the corpus, while 721 were discarded, which required 22 days and 11 hours of AMT work for a cost of 435.71 $. Including the work of the expert to prepare the gold data and manage the pipeline, the overall operation was a success and proved the usability and efficiency of crowdsourcing for building the multilingual entailment corpus.

Because of language variability, rule based semantic processing requires solving a large number of small inference problems, like for instance knowing that if *"X work as Y"* then it also means that *"X was hired as Y"*. But the evaluation of such rules is problematic because no reference dataset exists and extrinsic evaluation in the context of an applicative task will not necessarily produce evaluation of the rules since it is difficult to assess to which cause the application performance measured must be imputed to. So [Zeichner et al., 2012] experimented using crowdsourcing (CrowdFlower) to perform inference rule evaluation. To asses a particular rule application, one must answer 3 questions:

1. Is the left-hand side of the rule meaningful?
2. Is the right-hand side of the rule meaningful?
3. If yes was to answer to both previous questions, does the entailment holds?

There were 2 types of cascading tasks submitted to the annotators:

1. to appreciate the rule relevance, which corresponds to the first two previous questions
2. and to judge the entailment.

Depending on the answer produced by a task from the first type, the annotators working on the second type of task has either to validate the entailment or to ascribe the identified non-relevance of the rule to an erroneous answer provided for one the three original evaluation questions. Four inference algorithms were tested on a database of predicates extracted from ClueWeb09 web crawl2, where each extraction comprises a predicate and two arguments, providing four datasets. For each 5,000 extraction were sampled and for each dataset four rules common to all datasets were extracted which resulted in 20,000 rule applications, out of which 10,443 were discarded due to low CrowdFlower annotators confidence, a further 1,281 were flagged as meaningless left-hand side applications and another 1,012 as meaningless righ-hand side applications. Out of the remaining 8,264. rule applications that were passed on to the second type of task (entailment assessment), 5,555 were judged with a confidence high enough, 2,445 with positive entailment and 3,108 with negative entailment. In the experiment a total

of 6,567 application rules were annotated for a total cost of 1,000 $. Note that the second type of task required specific experimentation from the authors with the annotators in order to calibrate the communication of the task in order to improve the kappa performance. Th experiment was considered a success by the authors.

Statistical Modality Tagging was investigated by [Prabhakaran et al., 2012] Modality can be defined as a quantification over possible worlds or a speaker attitudes with respect to a proposition. Acquiring training data for building an automatic modality tagger is a difficult problem. In a pilot study, the author obtained and ran the modality tagger described in [Baker et al., 2010] on the English side of the Urdu-English LDC language pack. Using AMT, they estimated the precision of this type of approach at around 60%. They posted on AMT a set of randomly selected sentences (1997) that the tagger had labeled as not having the Want modality. Each sentence was checked by 3 annotators to decide whether it had the Want modality or not. Using majority rules on the annotations, 95 (4.76%) of the sentences were marked as validated Another set of 1993 sentences annotated by the tagger as not having the Want modality were posted on AMT, out of which 1238 were validated by the annotators. Hence, the authors decided to apply a simple tagger as a first pass, with positive examples subsequently hand-annotated through AMT. The simple tagger used a word spotting approach with a set of trigger words like “try”, “plan”, “aim”, “wish”, “want” etc. In the corpus, the number of sentences for each modality was limited to 50 for each trigger word in order to preserve linguistic variability. The AMT annotators were asked to check that the modality was not present in the sentence, otherwise they had to highlight the target of the modality. Each sentence was annotated by 3 persons. Only the output produced by adult annotator with an approval rating above 95% were considered if they had completed at least 50 tasks. They were paid 0.10 $ for each set of ten sentences. Only the annotation which gathered the approval of at least 2 annotators out of 3 on the modality and target were kept. Out of the resulting 1,008 examples, 674 had 2 annotators agreeing while 334 collected unanimous agreement. This work proved that it is possible to combine a high-recall simple tagger with crowdsourcing annotations to produce training data for a modality tagging.

[Lafourcade and Fort, 2014] propose to use the paradigm of GWAP (Game With A Purpose) to build lists of semantically-related terms, for instance needed to to deploy parental control systems on Internet. At the heart of the system is a GWAP that is used to build a semantic network where players collect credits when they provide answers similar to other players, with higher rewards for original contributions. The game played involve lexico-semantic relations like is-a, hyponym, characteristic, location, agent, patient, etc. In their paper the authors propose an algorithm to exploit the resulting semantic network to deploy a filtering service. The network produced for the oldest game on French contained in October 2013 approximately 300,000 terms, including 15,000 to 20,000 word usages and more than 6 million relations. The main game has been played more than 1.3 million times by more than 3,500 registered players. The experiment demonstrated the possibility to build and dynamically update a semantic network through crowd sourcing with a GWAP interface.

2.2.2 Opinion Mining & Sentiment Analysis

Also from 2009, the work of [Hsueh et al., 2009] report on the use of AMT to classifiy text snippets extracted from political blogs, first according to the the political candidate they concern and the polarity they express (Positive, Negative, Both, or Neutral) and second, according to whether they support or oppose the political candidate they are related to. The measured agreement among three expert annotator on the relevance to a particular candidate was 77.8% while overall agreement on the four-classes sentiment annotation was 70.4%. For the second classication step (support/oppose/neutral) the author report an agreement of 76.8%. In comparison, each snippet of a set of 100, was marked by 5 AMT annotators taken from a group of 25 annotators selected on the basis of an approval rate higher thatn 95% and paid 0.04 $ per annotation. Each snippet required on average 40s to be completed and the lower overall agreement measured on all four-class sentiment task was 35.3%. On the second classification subtask (support/oppose/neutral) they achieved a value of 47.2% of agreement, but they nevertheless managed to reach a relatively good level of agreement on the simpler tasks of determining whether a snippet is releveant to a particular political candidate (81.0%), whether the snippet is subjective or not (81.8%) and whether the snippet is positive versus negative (61.9%). But there exists a group of annotators that produces more noisy annotations than the other, judging against annotations produced by a majority vote, 20% of the AMT annotators have a noise level that exceeds 60% which disagree in 70% of the cases with the result of majority. The more the text snippet is ambiguous the lower is the agreement. In this paper 3 quality measures were found to be useful for selecting annotations: the noise level of annotator, the inherent ambiguity of the class labels and the informativeness of the annotated data.

A fine grained annotation task identifying word expressing a sentiment about a particular in-sentence target was experimented by [Sayeed et al., 2011] with CrowdFlower. This taks is in a simpler form (no identification of text spans required by the annotators), the task that will be used in the uComp evaluation campaign.The annotators had to classify as POSITIVE, NEGATIVE or NONE the preselected opinion words depending on their relation with a preselected target word located in the same sentence. The authors used the same techniques as [Hsueh et al., 2009] for discarding noisy annotations. The experiment had 200 tasks paid 0.04 $ per task, with three different annotators performing each task, for a total cost 60 $ and a time span of 24 hours to complete the job. In addition, 30 tasks were used to define ar gold standard which served to identify unreliable annotators (with less than 65% accuracy). This gold standard produced 117 words annotated as NONE, 35 as POSITIVE and 17 as NEGATIVE. Aggregation of the information produced by the annotators was done by majority voting (agreement above 50%) at a word level. There were 155 words with a majority consensus was reached (>50%). The authors of the paper determined 48 to have a particular opinion weight (POSITIVE or NEGATIVE). Only 22 annotators passed the Crowdflower quality control. Removing unreliable annotators based on the gold standard test had a remarkable effect on the F-measure and kappa values. the best kappa measure (0.65) was achieved when the 7 worse ranked annotators had been dropped. But highest precision (0.85) and accuracy (0.88) were

achieved after dropping the 10 worse annotators, note that in that case the kappa dropped to 0.54, because when going over 80% limit, performance start to decrease if more annotators were dropped because they were not numerous enough anymore for the voting mechanism to have a smoothing effect. This experiment showed that fine grained annotation is possible with crowdsourcing but requires extensive quality control procedures.

2.2.3 Named Entities

CrowdFlower is mentioned along AMT in [Finin et al., 2010], which describes an experiment about the annotation of Named Entities in Twitter (person, organization, location, or none) and the collection of judgements on the quality of "word clouds" for semantics and sentiment representation. The dataset was split into tasks containing 4 previously unlabeled tweets and 1 previously labeled tweet. On AMT, 251 tasks were submitted, each to be completed twice for an overall duration of 15 hours to complete the whole set of tasks at a total cost 27.61 $, which corresponds to 0.0275 $ per tweet. There were 42 AMT annotators mostly from the US and India and some from Australia. Most annotators performed a single task and one annotator did most of them. Inter-annotator agreement was checked using an algorithm akin to Google's PageRank ([Page et al., 1999]). The effectiveness of AMT annotators was judged inferior the to the one of the expert annotators but it was possible nevertheless to achieve the same results at a lower costs by by carefully combining annotations. The CrowdFlower experiment involved 30 tweets and each task had 3 tweets for a price of 0.05 $ to be done in 30s and representing a total cost of 2.19 $ (overhead included), and a rough pay of 2 $ per hour per annotator. The cloud comparison experiment was run on AMT, a task consisting in deciding which of the two word cloud presented to the annotator describes best the query topic. After selection with an average accuracy rating of at least 0.75 % on 7 questions, there were 8 AMT annotators selected and they did achieved a performance level of 61% of accuracy against gold data. The authors found particularly helpful the extra functionalities for managing the task and validating the annotators work provided by CrowdFlower over those of AMT.

[Sayeed et al., 2010] deploy crowdsourcing tasks to evaluate name entity recognition algorithms. They use AMT to assess the performance of an algorithm that identifies names of persons and organizations (ENAMEX NIST ACE-standard). For each task, each annotator was paid 0.05 $, summing up to 150.00 $ for 3 annotators. It took more than an estimated two person weeks to complete the work. They showed that crowdsourcing can provide reliable results and provide simple means of verifying algorithm performances, in a context where the aim is to reduce the rate of false positives.

The authors of [Higgins et al., 2010] describe how they used AMT to collect Arabic nicknames for completing exiting Named Entity lexicons. In addition, they experimented the effect that increasing the pay rate had on taks completion speed. On average, a pay of 0.03 $ per task yielded 9.8 names per day, increasing the pay to 0.05 $ made the number of collected names up

to 25 names per day and to reach the value of 100 names per day we had to go up to 0.25 $ per task. In another experiment they increased the price from 0.01 $ per task to 0.05 $, of course increasin the price shortened the collection time, but surprisingly enough, the publication of the higher paying task had also an accelerating effect on the lower paying task.

2.2.4 Machine Translation

Translation from Urdu to English was the theme of the AMT experiment reported in [Zaidan and Callison-Burch, 2011], the price paid was 0.10 $ per translation (approx. 0.005 $ per word). Four translations for each sentence were collected from different translators and a set of AMT task was created to cross validate the translations by other AMT annotators (ranking four translations at a time from best to worse or postediting). The task were paid on a basis of 0.10 $ for translating a sentence and 0.25 $ to edit a set of ten sentences, and 0.06 $ to rank a set of four translations. The overall costs were:

- translation cost: 716.80 $,
- editing cost: 447.50 $,
- Ranking cost: 134.40 $.

Adding Amazon's 10% fee resulted in a total less than 1,500 $ to more than 7,000 translations produce around 17,000 edited translations and rank 35,000 labels (since each ranking task involved judging 16 translations, in groups of four). Including the cost of the professional translation reference would add 1,000 $ to the total cost which now amounts to 2,500 $. The participation of the AMT translators was 52 for translation (138 sentences on average), 320 for editing (56 sentence on average) and 245 for ranking 245 Turkers (averaging 9.1 ranking task each, or 146 rank labels). The authors investigated cost reduction by eliminating the need for professional translation and decreasing the amount of edited translations. The first measure produced a significant drop of quality (BLEU score of 34.86), while the second measure greatly reduced the cost but managed to maintain good performance (BLEU score of 38.67). Then, the best strategy for translation seems to be

- for each source sentence produce several translations,
- rank the multiple translation,
- edit only the top ranked translations.

The work of [Hu et al., 2011] does not make use of AMT or CrowdFlower but is nevertheless interesting as it deploys two crowds of translators in a setup comparable with this two collaborative infrastructures. The experiment combined machine translation with human computation using two crowds of monolingual source (Haitian Creole) and target (English) speakers, for the WMT 2011 Haitian Creole to English translation task. The result showed that the combined approach translated 38% of the sentences well compared to Google Translate's 25%. The 4

Haitian Creole speakers were recruited from Haiti and did not speak English while the 26 English speakers were for 6 of them paid UMD undergraduates while the other 21 were volunteers of various background. Over 13 days, the total effort was 15 hours for Creole and 29 hours for English. The Haitian Creole sentence was first automatically translated into English and presented to the English speakers who could take one of the following actions:

1. mark a phrase in the candidate as an error,
2. suggest a new translation candidate translation,
3. change the ranking of the candidate translation.

New translation candidates were back translated into Haitian Creole and along with spans marked as errors which were projected back to identify the corresponding spans in the source sentence by means of word alignment. In turn, the Haitian Creole speakers could:

1. Rephrase the entire source sentence,
2. explain spans marked as errors,
3. basing their action on the back translation, change the ranking of the candidate translation.

Source speakers could document error spans either by rephrasing, either by annotating the spans with images or Web links (e.g. Wikipedia). The process was asynchronous for participants from both locations and the voting based best translation could be extracted at anytime.

[Post et al., 2012] applied AMT to the building of a collection of parallel corpora between English and six languages from the Indian subcontinent, low-resourced and under-studied which are difficult form machine translation: Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. The source documents in English were the top-100 most viewed wikipedia pages for from each language. Since the authors were not proefficient in all the six Indian languages, the decided to appreciate the quality of the AMT translator by comparing their production with the lexical translation of the original sentence by means of bilingual dictionaries bootstrapped for each language by means of another set of AMT tasks. In this other set of disctionary tasks, AMT translators had single words or very short phrases to translate while the validation was done with the wikipedia page titles whose translation can be assumed to be found in Wikipedia following the cross-lingual links. In the translation work, subsequent sentences from the original text were grouped by 10 to provide an AMT task, and translators had to provide a free-form translation. They were paid 0.70 $ per task. Decision to accept or reject the translation was done manually, by checking various factors like comparison to a lexical translation obtaine with the dictionaries developped for the project,the percentage of empty translations, the amount of time that the translator required to complete the task, his geographic location (self reported and identified through is IP address) and by measuring the distances between the various translation of the same source excerpt. Malayalam provided the highest through- put, generating half a million

words in just under a week. Some authors assess the cost of professional translation from Tamil to English at 0.30 $ per word while the cost measured in our experiment was less than 0.01 $ per word. Because no professional translation was available to the authors they could not use direct performance measures like BLEU to assess the variation of quality of the translations. Nevertheless, they designed another task in which original sentences were displayed along with four of its translations to another set of AMT translators for ranking the best. Each task was performed by 5 translators. Approximately 65% of the sentences had five votes cast on just 1 or 2 translations while 95% of the sentences had all the votes attributed to 1 to 3 sentences. This suggests that differences in translation quality existed but also that the translators took their did their assesment job seriously enough to report on the differences.

2.2.5 Information Retrieval/Extraction & Question Answering

The study of [Grady and Lease, 2010] is about human factor parametrization for a crowdsourcing tasks of relevance judgement in information retrieval. The author investigated, on AMT, the impact that the following four parameters had on the cost, time, and accuracy of the assessments:

1. providing the annotator with only a title for a query versus a detailled description,
2. different wording for the task title (specialized, i.e. “binary relevance judgement”, versus layman, i.e. yes/no),
3. the amount paid per task (0.01 $ versus 0.02 $),
4. and a bonus: (0 versus 0.02 $).

Assessment were done with document from the TREC TIPSTER collection of news articles (using gold standard data enabled easy computation of AMR annotators accuracy). Five batch evaluations were done, for each of four topics, five documents were assessed and 10 assessments were collected for each document. A total of 200 tasks were submitted to AMT for each batch, resulting in 1000 tasks for the five batches together. The length of the documents was between 162 words and 2129 words (including HTML tags and single-character tokens). For each task, the annotater had to perform a single binary relevance judgment linking a query and a document. There were 149 annotators who participated, some of them to the 5 batches. In batches 2 and 3 one variable was modified with respect to batch 1, while in Batches 4 and 5 it was against batch 3 configuration that one variable was modified. In batch 5, 23 bonuses were given for a total cost of 0.46 $. Statistical significance was measured via a twotailed unpaired t-test. The only significant outcomes observed were increase in comment length and number of comments for higher-paying or bonus batches. The hihgest accuracy, 70.5% was reached with batch 3, which used a title query and a simple yes/no response. The use of description query did not entailed an accurary improvement. The fastest task completion, 72s, was measured for batch 4 while the average time completion per task across the five batches was 63s. The number of unique annotators per batch varied between 64 to 72 for batches 1 to 4 but fell to 38 for batch 5, probably because of the bonus incentive, annotators in this last batch tended to complete

more tasks better accuracy (3.37 documents correctly annotated as compared to 2.10-2.20 for batches 1 to 3 and 1.85 for batch 4). But bonus attribution requires expert human supervision and the question remains of the gain versus expert time spent for the operation to become beneficial. From the words of the authors, this experiment provided largely inconclusive results.

The following work was realized with the help of a web application game to collect crowdsourced input. After qualitatively examining how humans perform incremental classification [Boyd-Graber et al., 2012] have shown in their article how crowdsourced knowledge of a human's incremental classification process improves state-of-the-art rapacious classification. Then they built a Bayesian models that embedded in a Markov decision process to replicate the improved classification process and develop new hierar- chical models combining local and thematic content to better capture the underlying content. The corpus used was made of 37,225 quiz bowl questions with 25,498 distinct labels from 121 tournaments (between 1999 and 2010). The authors created a web application to simulates the experience of playing quiz bowl where text is incrementally revealed until the user decides to answer. The answer is judged with a string matching algorihtm. More than 7000 questions were answered in the first day, and over 43000 questions were answered in two weeks by 461 users.

2.2.6 Language Communication Enhancement/Validation

It is to palliate the lack of common reference corpus for evaluating grammatical error detection that [Madnani et al., 2011] decided to use crowdsourcing. For their experiment, the authors studied with AMT the presence of extraneous preposition in a corpus of students writing for a test of English as a foregin language. In this experiment 75 sentences were used as gold standard built by 3 experts and the remaining 923 sentences were annotated by 20 annotators located in the USA, whithin one day. Using 3 annotators per judgement with a majority vote yields an agreement with any one of annotator of 0.87 on average which corresponds to a kappa of 0.76. The extraneous preposition annotation costed 325 $. Further experiments were done with a gold standard of 20 sentences obtained with CrowdFlower, in particular to propose new evaluation measures, derived from precision and recall by weighting the evaluation of annotations items depending on the proportion of agreement for this item by the annotators. The authors found out that the weighted measures are more stable and contrary to regular precision and recall they display less a tendency to overestimate the performance of the system under evaluation. Similarly, for comparing two systems they propose to replace precision-agreement estimation by a kappa-agreement measure

Augmented and Alternative Communication (AAC) devices enable users with communication disabilities to participate in everyday conversations. For designing such devices, the elaboration of language models representing as best as possible the style of the users' intended communications is essential. Since collection of geniune AAC material for designing such device is quite difficult, the authors of [Vertanen and Kristensson, 2011] decided to use crowdsourcing

to create a corpus of fictional AAC messages. In their paper, they show that the messages they produced a better model of AAC communication than the material generally used so far from telephone conversations or newswire text. Two types of tasks were proposed:

1. the AMT workers tried to imagine how he would used an AAC device to communicate if he was subject to a communication impairement,
2. AMT annotators were asked to assess the plausibility of communications produced by the previous type of task.

Task of the second type gathered more easily AMT participants that tasks of the first type. The average task completion time was shorter (24s on average) for the second type of task, as opposed to an average of 36s for the tasks of the first type. During the experiment, the pay for the the second type of task was reduced from 0.04 $ (the price paid for both types of tasks) to 0.02 $ whithout loss of performance. For comparisons the authors trained language model with material of different origins (Wikipedia discussions, Usenet, Switchboard, newspapers, Twitter...) and compared the performance with models trained on the crowdsourced data. Compared to a model trained only on Switchboard, their best model reduced perplexity by 60-82% relative on three AAC-like test sets, which represents a potential keystroke saving of 5–11% on a predictive keyboard interface.

2.2.7 Real Time Dialog System

The work of [Bessho et al., 2012] addresses the creation of utterance/reply pairs for Japanese dialog system design. Here the wisdom of the crowd is not provided by a crowdsourcing framework but by Twitter. For each user input, the system will extract from the utterance-pair database the pair for which the tweet is mbost similar to the input part of the pair and the system response is provided by the output part of the pair. The utterance pair database was built using a corpus of 1.2 million utterance-pairs from Twitter which were written in Japanese, contemporary, and had a in-reply-to field. The author propose to integrate into the dialog system a "real-time crowdsourcing" functionality to handle the cases where the system cannot provide an adequate answer (similarity distance is below a certain threshold between processed utterance an input parts of the utterance pairs stored in the database). The original user input is recast into a tweet from the dialog system chatbot and if a crowd member responds before a certain delay the crowd answer is used as reply by the system. For evaluation, 90 user input examples were selected and 20 utterance-pairs were extracted from the database retrieved from Twitter for each per user input, totaling to 1,800 of triples (user input and utterance pair). Thirty subjects evaluated naturalness and versatility of the responses (600 triples each). Various scoring functions were investigated by means of a ROC curve representation (true positives versus false positives). The area under the curve (AUC) was used to measure the classifiers performance. A random classifier has an AUC of 0.5, and ideal classifier has an AUC of 1.0. The scoring function selected was the one with the best performance, here an AUC of 0.803.

2.2.8 Conclusion

The conclusion that we draw from this survey is that crowdsourcing or similar approaches like collaborative ones mediated by Internet or GWAPs ([Lafourcade and Fort, 2014]) offer the possiblity to develop language ressources for which human language processing functionality is required at a cost much lower than with classsical means of production, for various kind of tasks, in particular task like semantics or translation. This is so because the people employed in this kind of infrastructure are recruted in large numbers, potentially from all over the planet and do not need to have a particular expertise except of being proefficient in (most often) one or several languages. If the number can compensate the quality of the individual annotations by cross-validation, the measure that a task proposer has to deploy to ensure a minimum of quality require either gold standard data in sufficient number or the involvement of an expert, in addition to designing specific procedures to prevent cheating. Despite the extra cost incurred by the measures deployed to ensure quality of the resource produced by crowdsourcing, if the feasability study has been properly done the resulting cost is much lower than with traditional means. But crowdsourcing infrastructure evolve rapidly and crowds of participants adapts rapidly to new tasks which means that quality measure need to be revised frequently.

There is a question that is rarely raised in the litterature, nevertheless addressed by [Fort et al., 2014] and also the following presentations available on the Web [Lease, 2013a], [Lease, 2013b], [Larson, 2013], it is the question of ethics about the minimal return pay somebody working full time in a crowdsourcing infrastructure is able to achieve. As a corrollary, one cannot but consider the question that the new infrastructures for crowdsourcing raises about social tax evasion.

2.3 7 language lexica Experiment

The first part of Deliverable D5.2 consists in the 7 language lexica extracted using as baseline the pointwise mutual information (PMI) measure ([Manning and Schütze, 2002]) collected using climate change linguistic markers obtained by machine translation from the set of patterns developped for English, German and French [Scharl, 2014].

D5.2 V1# of entries	D5.2 V2	Language
12,137	0	en
9,394	0	es
6,129	0	it
5,606	0	pt
5,294	9,930	fr
4,336	0	de
760	0	ru

Table 13: Number of entries for the 7 languages in first version of D5.2

Since the amount of data was too small, a second corpus extraction was done with a wider source set (less focused on institutional sources), initially on French only, which resulted in a 7,000 Tweets corpus, out of which a 9,930 entry lexicon was extracted using PMI, then further lemmatized by alignment with the content of the LEFF lexicon [Sagot, 2010].

2.3.1 The HC experiment with CrowdFlower

Concerning the comparison between automatic extraction of lexica and HC computations, there is an ongoing experiment running on Crowdflower with the French lexicon as input. The parametrization of the experiment is as follows:

- Each annotation task consists in a set of 8 lexical entries (units in Crowdflower terminology, note that the number of units per task advised by Crowdflower was 5) for which the annotator must provide answers to two questions:
  1. Is the term related to an expression of opinion/sentiment/emotion (yes/no)?
  2. Which among the following 20 categories ([Fraisse and Paroubek, 2013]) of opinion/sentiment/emotion is the most appropriate to describe the meaning of the term?
- the price set of the completion of a task is 0.07 €(not that the price advised by Crowdflower was 0.10€),
- the total cost of the experiment is 200 €
- the experiment was started on Monday, July 21th 2014,

The experiment was submitted directly to Crowdflower, we did not test the GATE API yet.

Although the task is not yet completed, we already have gained a useful information concerning the potential use of services like Crowdflower with respect to difficulty to provide a task description for relatively complex classification schemes, since the amount of information that one can provide to the *taskers* is relatively limited.

2.3.1.1 Job design

As described in Figure 1 data consists of 9939 terms.

As shown un Figure 2 unit consists in the 4 questions:

- This term may be used to express opinion, sentiment or emotion ?
- This term may be a trigger of an opinion, sentiment or emotion ?
- What is the connotation of this term ?
- what type of opinion, sentiment or emotion does this term express ?

Add More Data | Split Column | Convert Uploaded Test Questions

Unit ID	State	Judgments	Agreement	scratchage
506076166	Golden	423		indignation
506076189	Judgable	0		rationaliser
506076195	Judgable	1		ignorance
506076196	Judgable	0		artificieux
506076212	Judgable	0		implacable
506076221	Judgable	0		déserter
506076222	Judgable	0		agressive
506076231	Judgable	0		givrure
506076232	Judgable	0		crêpage
506076233	Judgable	0		réembarcation
506076234	Judgable	0		pas-de-porte
506076235	Finalized	4	0.875	polychète
506076236	Judgable	4		polyviser
506076237	Finalized	5	0.55	au mieux
506076238	Finalized	5	0.9	cartouchière
506076239	Judgable	1		désindemniser
506076240	Judgable	1		dualisation
506076241	Judgable	0		guiper
506076242	Judgable	0		napoléon
506076243	Judgable	0		esquintant
506076244	Judgable	0		caviarder
506076245	Judgable	0		héliciculteur
506076246	Finalized	4	0.9375	talonnière
506076247	Finalized	5	0.85	retour
506076248	Finalized	5	0.6	criminalisation

Page 1 of 398 | Displaying units 1 - 25 of 9939

Figure 1: Data sample

2.3.1.2 Quality management

Concerning the quality control, we defined one test question (Figure 3). However, in order to have as many contributors as possible, we selected the *performance level 1* (the lowest level proposed by CrowdFlower) for French (Figure **??**). For the CrowdFlower job, we put 10 units per task that paid 0.06 €(Figure **??**) for a task to be completed whithin an expiration delay of 30 minutes.

2.3.1.3 Results

The job was running for 3 months and as shown in Figure 4, only 9.1% of the job was completed. This results can be explained first by the complexity of the task. In fact the fine-grained classification task is more complicate than a simple binary classification task and this is true even if the contributors are native speakers. The second raison, for this low score is the payment of the task. In fact, contributors are less motivated when the task is underpaid (0.06 €instead of 0.10 €that was recommended by CrowdFlower).

In the Figure 5, each bar represents a contributor and the number of judgments they have submitted for this job. Contributors who have a low trust score and have submitted a significantly larger amount of judgments than other contributors are likely scammers. So, based on

ignorance

Ce terme peut-il être utilisé pour exprimer une opinion, un sentiment ou une émotion ?

- ◯ Oui
- ◯ Non

Ce terme peut-il déclencher une opinion, une sentiment ou une émotion ?

- ◯ Oui
- ◯ Non

Ce terme a t-il une connotation

- ◯ Positive
- ◯ Negative
- ◯ Neutre

Quelle est l'opinion, le sentiment ou l'émotion que ce terme peut exprimer ou déclencher ?

- ☐ AMOUR
- ☐ APPAISEMENT
- ☐ DERANGEMENT
- ☐ ENNUI
- ☐ PLAISR
- ☐ DEPLAISIR
- ☐ TRISTESSE
- ☐ PEUR
- ☐ COLÈRE
- ☐ SURPRISE POSITIVE
- ☐ SURPRISE NEGATIVE
- ☐ SATISFACTION
- ☐ INSATISFACTION
- ☐ VALORISATION
- ☐ DEVALORISATION
- ☐ ACCORD
- ☐ DESACCORD
- ☐ DÈGOÛT
- ☐ AUCUNE (TERME NEUTRE)

Figure 2: Unit description

Questions Settings Stats

Filter judgments: All Quiz mode Work mode

No filters

Reports Create More

1 total Test Question

ID	Judgments	% Missed	% Contested	Last updated	Enabled	Actions
506076166	423			3 months ago by fraisse@limsi.fr	On	Show Details

Figure 3: Test question

the gold data set, their work will be rejected. Only the top 100 contributors are displayed in this graph.

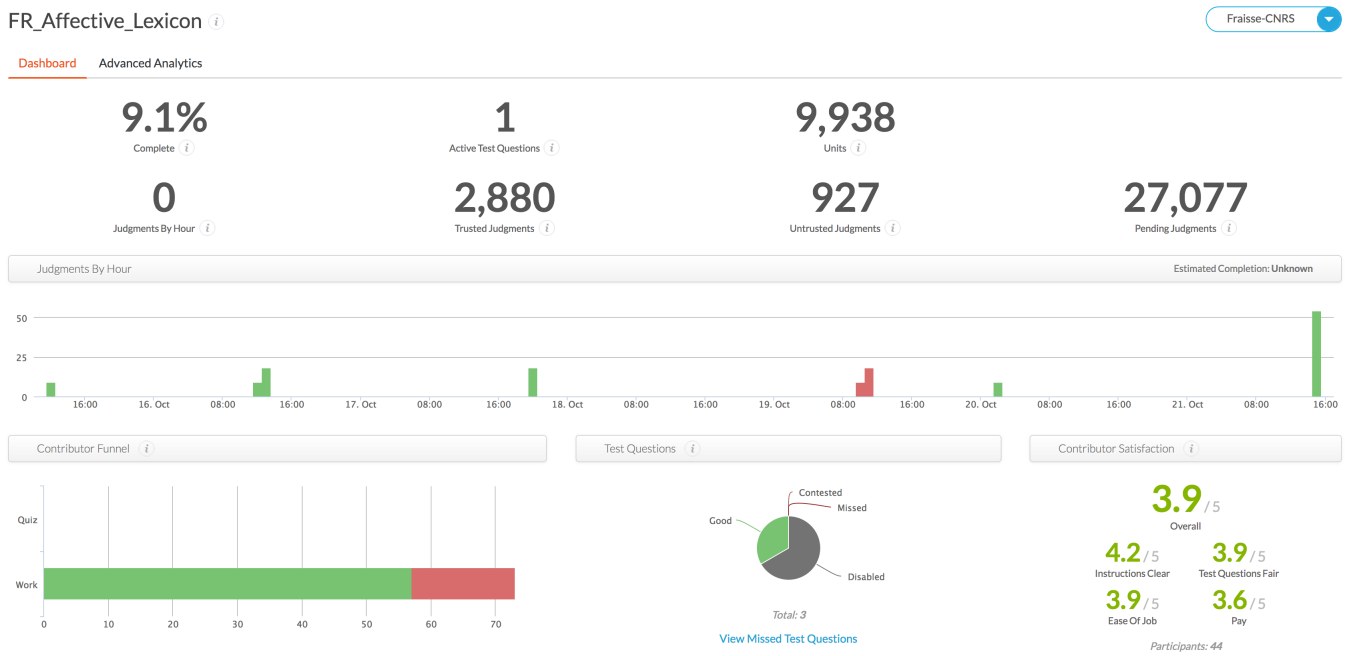

Figure 4: Dashboard of the Workflow job concerning the validation of the french affective lexicon

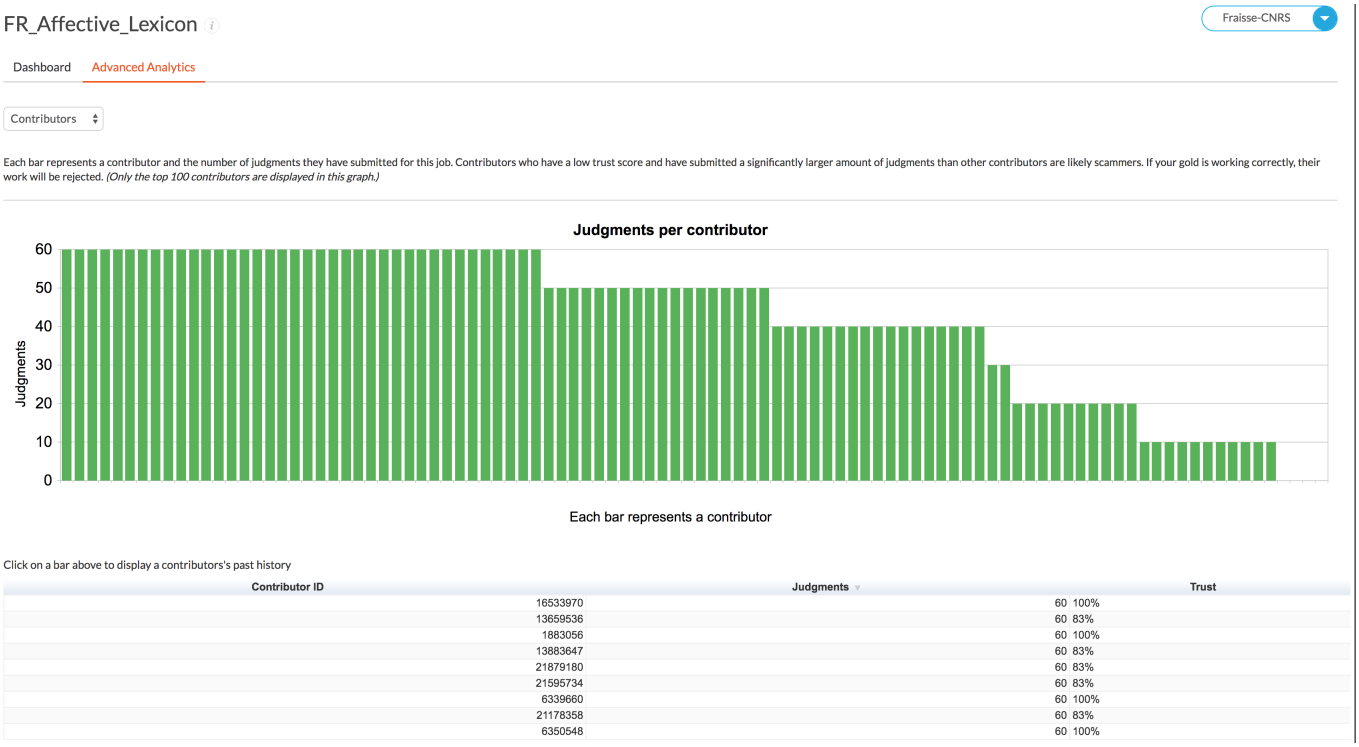

Figure 5: Judgments per contributor

3 Evaluation Campaign

3.1 Aim

T5.3 aims at organizing an open evaluation campaign on affective document analysis (identifying source, target and affect expressions) with a corpus-based quantitative black-box evaluation methodology in the climate change domain, for French and German. The HC-based workflow developed in WP2 will be used first to build and validate the gold standard, second to propose input resources to the participants (ontology elements of WP4, affect lexicons, etc.). Thus a participant will be able to either use their own resources, or build on the proposed input resources to improve their system (or both, in which case we will be able to assess the impact of the resource choice on system performance). The campaign will also be an occasion to test replacing a static comparison against a gold standard by a dynamic assessment of the participants' data through HC.

3.2 Data

A public call for tender has been published in February 2014 by LIMSI-CNRS (PUMA Nbr. 43826) for providing reference annotations on microblog textual data (Tweets), for an amount of 4,200,000 signs (or 30,000 messages) equally spread between French and German (15,000 of

each). ELDA, the Evaluations and Language resources Distribution Agency (located 9, rue des Cordelières, 75013 Paris) was the only institution to make an offer and was selected.

So far, a the French subcorpus was extracted by MODUL from the data collected with the *Media Watch on Climate Change* demonstrator. A query with a list of regular expressions [Scharl, 2014] resulted in 3,433 tweets, 112 YouTube videos, 77 Facebook messages and 11 Google+ postings for the period 26 March to 26 May 2014. The first exploration of the corpus showed a relatively high proportionof retweets or purely informative tweets. LIMSI using the same set of regular expression queries did an new extraction from Twitter (July 2014) on French which resulted in a corpus with a relatively higher proportion of subjective messages, containing 7,000 posts. Preliminary tests for the deploying the annotation framework have been done in collaboration with ELDA in July 2014 and production of the annotation of the French reference corpus is planned to start in full during August 2014.

3.3 A model of language data annotation

Comparing the HC framework against the classical annotation approach for producing evaluation ressources calls for some automatic support for what concerns the mapping from the classical linguistic annotation scheme, designed for expert annotators, onto a smaller grain annotation procedure suitable for crowdsourcing or GWAPS users. To this end we first provide a task generic annotation formalisation which we will use in a second step to draw the specification of a data converter that will automatically prepare a corpus for crowd annotation from a corpus prepared for classical annotation, allowing the task designer to specify relevant parameters.

3.3.1 Annotation representation

In what follows, we call "control task", the information processing task for which we want to assess the performance of some computer system, for instance if we look at Information Retrieval (IR), given a set of documents and a query, the control task consists in identifying which documents are relevant with respect to the given query [Cleverdon, 1960]. In fact, for any objective evaluation task, the systems under test are asked to output a symbolic represention in function of the objects and relations that they have found to be present in the input representation (the test data). The control task may be limited to identifying only the objects present in the test set, *e.g.* POS tagging [Paroubek, 2007] for which the output representation is made of the word boundaries and their class label. Sometimes the boundaries of the objects present in the test set are given and the systems need only to identify the class they belong to (*e.g.* IR or Word Sense Disambiguation [Edmonds and Kilgarriff, 2002]). On the other hand, some control tasks are much more complex and require identify objects, primitive relations holding between objects and also higher level relations holding between primitive relations, like in parsing [de la Clergerie et al., 2008], anaphoric resolution [Vilain et al., 1995] or image recognition [Unnikrishnan et al., 2007]. For tasks like machine translation, we are in general only interested in the final result of the transformation of the objects and relations that the system under test has identified in the input data.

In the most general case, a control task can be seen as a process that links together test data units resulting from a segmentation process and annotation symbols possibly organized in several layers. Assuming that the test data is the result of a segmentation process of an input medium (character stream, speech signal, pixel array, etc.) represented by $S = \{s_i/0 \leq i \leq N \in \mathbb{N}\}$, and the set of annotation labels by A, the m layers of relations graphs R resulting from the annotation process can be expressed as follows[2]:

$$\begin{aligned} R = & \bigcup_{j=1}^{m} R_j, \quad m \in \mathbb{N} \\ R_1 = & \bigcup_{k=1}^{q} \{r_l/l \in \mathbb{N}, r_l \subset \mathcal{P}(S^k \times A)\} \\ R_i = & \bigcup_{k=1}^{u} \{r \subset \mathcal{P}((S \cup R_{x_1}) \times (S \cup R_{x_2}) \cdots \times (S \cup R_{x_k}) \times A), 1 \leq x_k < i\} \end{aligned} \tag{1}$$

R_1 represents the first layer of annotation of the test data (i.e. the set of all relations of any arity between initial segmentation units) and the R_i represent the successive layers of annotations that may reference annotations from any previously existing layer down to the initial data segments themselves (see Figure 6).

Note that with respect to the *annotation graph* model from LDC [Bird and Liberman, 2000], which directly link annotations to events from the various linear input streams that they decorate, we encode in our model the (potentially recursive) structure of the annotations. Such information is important in our opinion when comparing annotation schemes to take into account their relative structural complexity. Our representation is very much like the one proposed in [Roth and Sammons, 2008] except that we adopt a slightly more general point of view by abstracting any implementation detail like annotations identifiers, types, constituents etc. to simply a kind of relation.

```
(S (NP-SBJ I) (VP consider (S (NP-SBJ Kris) (NP-PRD a fool))))
```

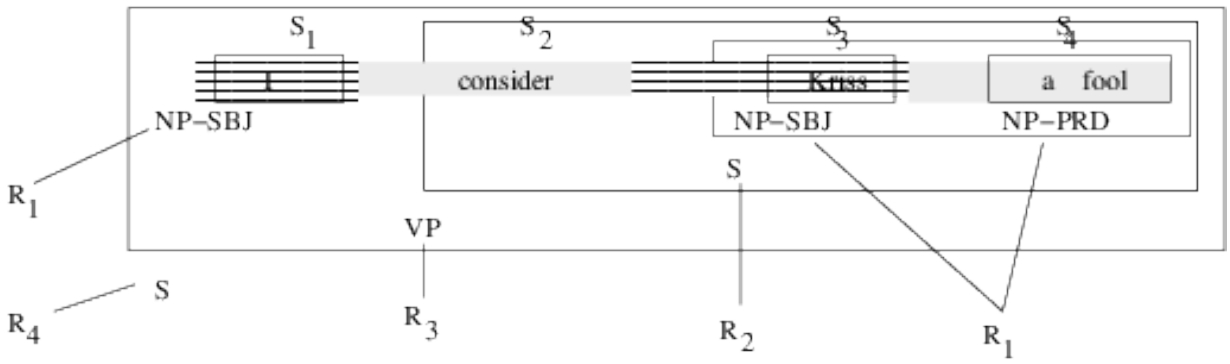

Figure 6: S_i and R_i for a PennTreebank syntactic annotation sample.

In the next section, we give examples of the instanciation of S, R_i and A for a selection of well known evaluation campaigns.

[2]In formula 1, $\mathcal{P}(x)$ is the set of all subsets of x.

In the literature, computer traces resulting from performing the control task are called *hypothesis* (annotations) and the human ones: *gold standard* or *reference* (annotations). The objective evaluation result is obtained by comparing the human and computer traces produced in response to the test data, *i.e.* comparing the set of relations identified by the computer (H) with the one identified by humans (G). When the evaluation is quantitative, the result is obtained by computing some *measure* defined over the two sets of relations. Here it is important to point out that the result of evaluation, which is a measure $R \times R \rightarrow \mathbb{R}$, is a function of T the test data, whose role in linking reference and hypothesis data is essential for the computation of the evaluation result. Even more so, when the system under test uses relations R' which have different semantics from the reference ones (but nevertheless mappable to them), or when it modifies the input data because of noise, data corruption or specific normalization, or when it uses a segmentation function S' different from the reference one. With N the "noise" function, the hypothesis is then better described as :

$$\begin{aligned} H = & \bigcup_{j=1}^{n} H_j, n \in \mathbb{N} \\ H_1 = & \bigcup_{k=1}^{q} \{S' : T \rightarrow \mathcal{P}(T), S' \neq S, A' \neq A, N : T \rightarrow T, r_l \subset \mathcal{P}(S'(N(T)))^k \times A')\} \\ H_i = & \bigcup_{k=1}^{u} \{r \subset \mathcal{P}((S'(N(T)) \cup H_{x_1}) \cdots \times (S'(N(T)) \cup H_{x_k}) \times A'), x_k < i\} \end{aligned} \tag{2}$$

In addition to the mapping μ from relation annotation labels A' to A, which in general is provided by the participating system, one must then be able to find a "reasonable" mapping M between $S(T)$ and $S'(N(T))$ to be able to compute an evaluation result. By "reasonable", we mean a mapping that maximizes the global similarity between the reference and hypothesis versions of the annotated material with respect to a particular similarity function σ. This is what is done for instance when one uses dynamic programming to find the mapping that minimizes the edit distance between two slightly different versions of the same text to compare their POS tag annotations [Paroubek et al., 1998].

$$M = \underset{m}{argmax} \sum_{h,g} \sigma(m(h), g), \tag{3}$$

$$m \in \mathcal{P}(S'(N(T))) \rightarrow \mathcal{P}(S(T)), \sigma : \mathcal{P}(S(T)) \times \mathcal{P}(S(T)) \rightarrow [0, 1]$$

Most of the time, an evaluation campaign will define several measures in conjunction and use the vector space corresponding to the measurement tuples to represent the performance of each system as a point in the n-dimensional Euclidean space. Their relative position is then characterized by their distance, justifying the use of the term *metric* instead of measure.

3.4 Applying our model to real evaluations

In this section, we use the model proposed in section **??** to represent well known evaluation protocols from different domains of natural language processing [Paroubek et al., 2007].

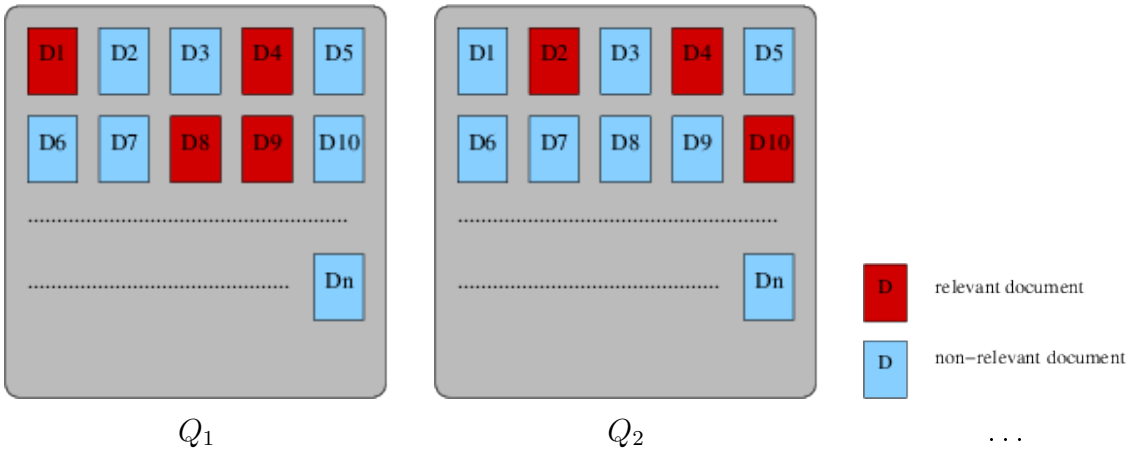

Figure 7: Classification of the set of documents among two parts (relevant and non relevant) for each query: for Q_1, D1, D4, D8 and D9 are relevant; for Q_2, D2, D4 and D10 are relevant.

3.4.1 Classification

The first type of applications that we identified is the general domain of classification and relation extraction. In this kind of task, the purpose is to segment a data set in order to highlight parts of this set that belong to specific classes (predefined or not), and possibly to provide relations existing between these segments.

We take the examples of information retrieval, named entity recognition, temporal annotation and parsing.

Information retrieval Classical information retrieval aims at finding full documents that are relevant to a given query Q. The document collection C is the input data (the retrieval unit is the entire document level). Here we consider a simplified instance of the general model presented in the previous section, in a sense that the segmentation of the test data into units to be annotated is provided, it is made of the documents themselves, see Figure 7.

This is a classification task, since the aim is to produce a partition of the collection, between relevant and non-relevant documents, with respect to the query. In practice the evaluation data contains several queries, but since in general they are considered independent of each other, the evaluation resolves to a series of single query evaluation (see Figure 7).

In other words, variables introduced in Section **??** are instanciated in the following way:

T : the set of documents

$S(T)$: the structuration corresponds to the existing document boundaries, $\forall t \in T, S(t) = \{t\}$

A : a set of two labels: *relevant for the query* or *not relevant for the query*

$R = R_1$: a singleton made of one unary relation that tag the relevance of each documents.

Figure 8: Classification of the sets of characters considering the named entity types (here, organizations, locations, dates, none).

Named entity recognition The following steps can be identified concerning named entity recognition:

1. Identification: finding which data units of the test set need to be annotated.
2. Categorization: finding the appropriate relation to annotate a data unit from the test set, *e.g.* tagging word sequences with labels for locations, persons, organisations, etc.

A normalization step can be added, as for example at the Temporal Expression Recognition and Normalization (TERN) Task of EVALITA [Magnini et al., 2008], where temporal expression should be associated with a universal representation of the expression. All NE types can be concerned by this normalization, for example person names, since they exhibit often many variations in their realization: "Barack Obama", "B. Obama", "President Obama", "Barack H. Obama".

Note that for named entity recognition, the segmentation function of the test data into elementary units is generally not provided by the evaluation organizers. This is not the case for the following example: TempEval.

T : a document, seen as a stream of words or characters
$S(T)$: the segmentation of NE types, at character or word level
A : the set of NE class labels
$R = R_1$: a singleton holding the unary relation linking the NE to its class label.

Temporal annotation Temporal annotation as defined by TempEval evaluation campaign [Verhagen et consists in the following: given a set of test texts for which sentence boundaries are annotated, as well as all temporal expressions and events in texts, the control task goal is to link events to other events, or events to time expressions (see Figure 10).

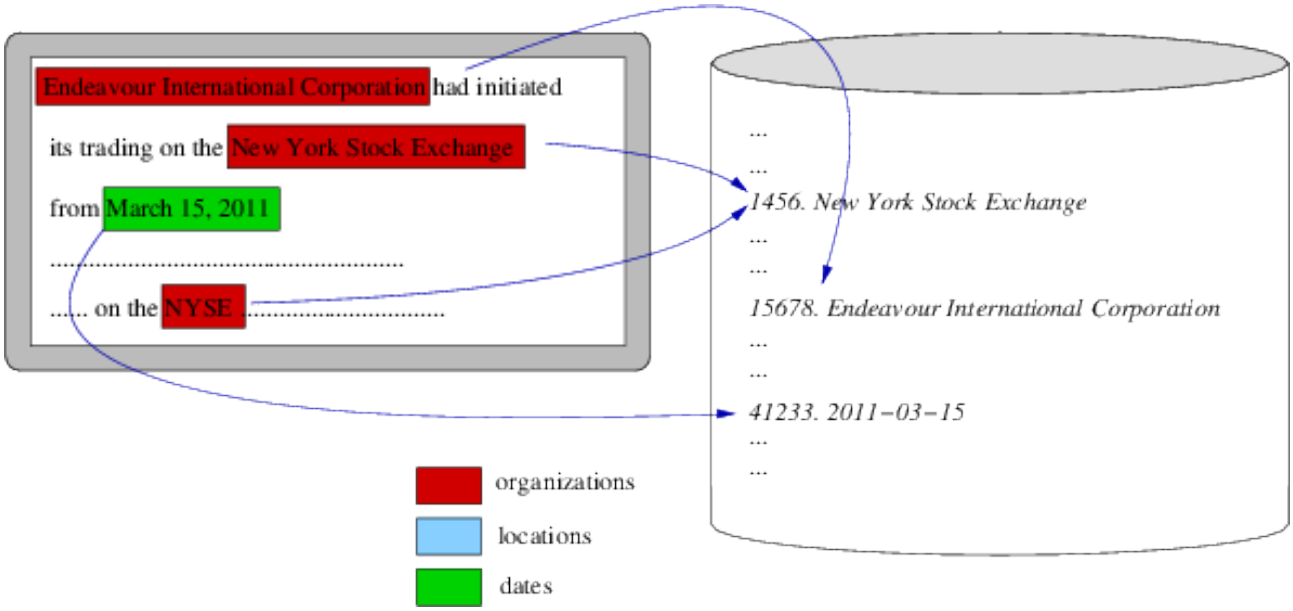

Figure 9: Classification of the sets of characters considering the named entity types and relations to normalized entities in a separate knowledge base.

T	:	a document, seen as a stream of words or characters.
$S(T)$	:	the segmentation into temporal expressions, signals, and events.
A	:	the set of temporal expression signal and event class labels, as well as temporal relations labels.
$R = R_1$	:	1/ the relation that links a temporal expression, signal or event to its class label, *e.g. kipnapped* is an event. 2/ plus all the labeled time relations between the temporal elements *e.g. kidnapped* is **before** *rescued.*

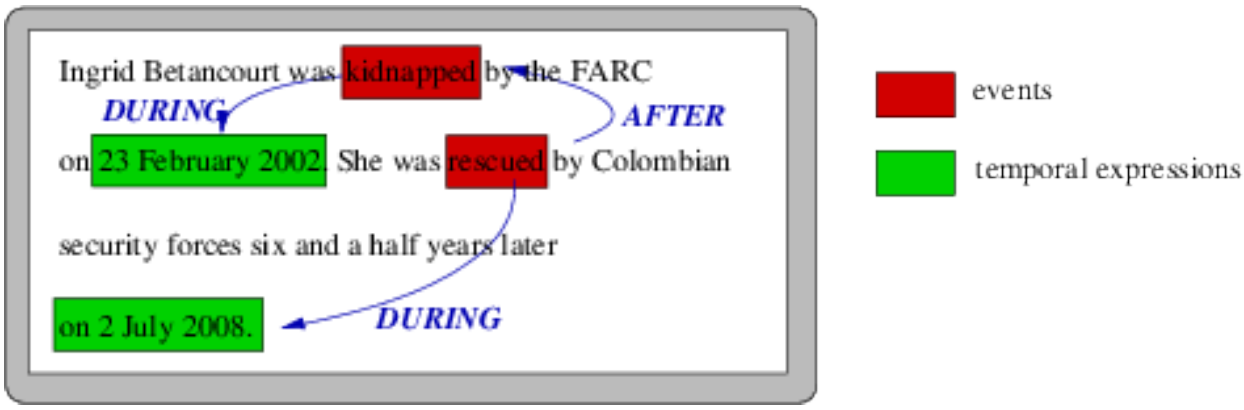

Figure 10: Temporal annotation.

Parsing The aim of automatic parsing is to provide a complete/partial structural analysis of a sentence expressed in terms of:

- chunks, sequences of words with some syntactic meaning,
- constituents, sequences of words which function as a single units within a hierarchical structure,
- dependencies, relations linking a particular word (the head) and one of its dependents,
- links, relation between pairs of words whithout necessarily referring to a tree hierarchy,
- grammatical relations, i.e. relation/head/dependent tuples [Watson et al., 2005],
- derivation/derived tree [Schmitz and Le Roux, 2008] describing the construction of the syntactic parse tree ,
- etc.

Since theories and annotation schemes are quite numerous and diverse in parsing, we present here only a few annotation schemes which have been used for evaluation: the PennTreebank [Marcus et al., 1993] for constituent analysis of English and PASSAGE [Vilnat et al., 2010] for chunks and grammatical relations in French. The PennTreebank example (see Figure 6 on page 23) is the first example of annotation scheme in this article which exhibits both relations between annotated elements (words) and their class label (e.g. the relation between NP-SBJ and "I"), as well as relations between annotations themselves (e.g. the toplevel relation between S and the constituents NP-SBJ and VP).

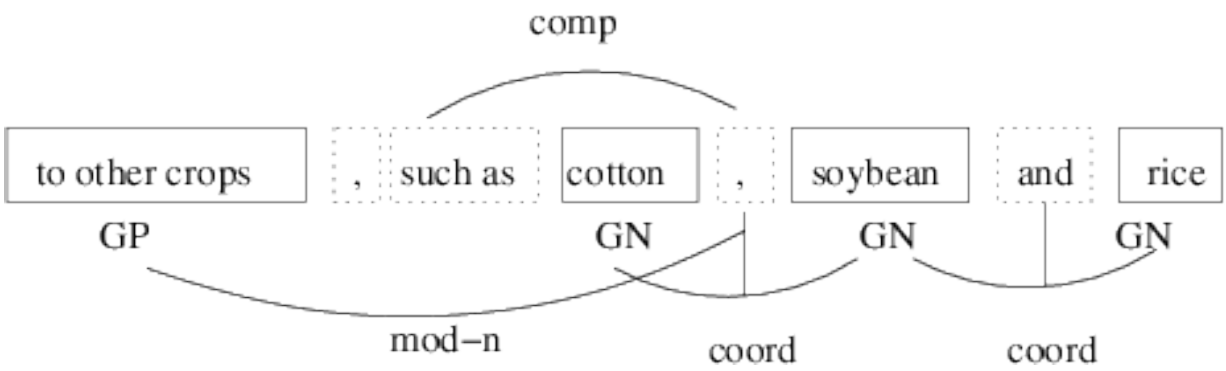

Figure 11: Example of PASSAGE annotation

The Penn Treebank constituent annotation model:

T : a documents seen as a stream of characters

$S(T)$: the segmentation into words

A : the set of constituent labels

R_1 : relations between words and their deepest layer of constituent label

R_i : relations between words and constituent labels, or between constituent labels of deeper levels

R : $\bigcup_{j=1}^{m} R_j, m \in \mathbb{N}$

The PASSAGE annotation scheme has only one layer of non-recursive syntactic chunks and grammatical relations defined between words and/or chunks (cf Figure 11). Element of comparison between the passage annotation scheme and PARC, SD and GR, three other syntactic annotation schemes used for English parsing evaluation are provided in [Paroubek et al., 2009].

The PASSAGE annotation model:

T : a documents seen as a stream of tokens

$S(T)$: the segmentation into words

A : the set of chunk and relation labels

R_1 : the relations between words and their chunk labels, or relations for which at least one argument is word (e.g. coordinating relation for whose coodinating conjunction argument is always a single word not included in any chunk, see Figure 11)

R_2 : the relations linking chunks only

R : $R_1 \cup R_2$

3.4.2 Transduction

Lastly, a very different type of applications is the set of applications producing an output that is not an enrichment (or annotation) of an existing test set, but a new object obtained by transformation from or in response to another object. Examples of transduction applications are: machine translation, speech synthesis, automatic summarization, language generation [Koller et al., 2010] or machine dialogue.

For all these examples, T is a document seen as a sequence of characters to be either translated, synthetized, summarized, etc. $S(T)$ is the existing segmentation into language units, while A is the result of the operation: the translation of a language unit into the target language, the synthesis of a language unit into sound generation instructions, etc. R is the set of links between language units in the test set and elements from A.

T : a documents seen as a stream of tokens

$S(T)$: the segmentation into transduction source units

A : the corresponding transduction target units labels

$R = R_1$: relations between source and target units

Note that we can consider multilingual alignment tasks [Chiao et al., 2006] to be degenerate cases of transduction task, where the target labels are provided as input data and the systems under test need only to identify the relation between source and target units.

3.5 Expert Annotation Guidelines

Opinion, Sentiment and Emotion annotations are applied on microblogs text message (maximum lenght of 140 characters). The annotations scheme was inspired from the PASSAGE evaluation campaign of syntactic parser of French for the structure to which a specific semantics for opinion, sentiment and emotion annotation were added. With the previous fromalism, the annotation structure is as follows :

T : a documents seen as a list of tokens
$S(T)$: the segmentation into words
A : the set of chunk and relation labels
R_1 : the relations between words and their chunk (group) labels, or relations for which at least one argument is word
R_2 : the relations linking chunks only
R : $R_1 \cup R_2$

In total the uComp annotation scheme holds 7 main types of groups (chunks) which resolve to 27 possible terminal labels (in the previous model $|A| = 27$).

1. HOLDER
2. Opinion/Sentiment/Emotion Expression (OSEE), which is further refined into different sub-categories, 21 of fine semantic grain split into 7 grained positive, 11 negative, one neutral for instructions or demands and 3 of coarse semantic grain: positive, neutral and negative, totaling in final 24 possible annotation labels for groups.
3. TARGET
4. NEGATION
5. MODIFIER
6. RECIPIENT

and 5 relations ($|R| = 5$):

1. SAY, that links an HOLDER to its OSEE.
2. ABOUT, which relates an OSEE to its TARGET.
3. MOD, connecting a modifier, like an adjective or an advert to OSEE.
4. NEG, associating a negation marker to an OSEE.

5. RECEPTEUR, that links the OSEE to its RECIPIENT when the OSEE if of subtype instruction or demand.

.

3.5.1 Groups

3.5.1.1 Holder

La source est constituée du groupe de mots qui référence l'auteur de l'expression d'opinion/sentiment/émoti (OSEE: Opinion Sentiment Emotion Expression).<p> The holder is the group of words, which refers to the author of the opinion, sentiment and emotion expression. If there is no explicit mention of the holder, we suppose that the writer of the text is the holder of the OSEE expression and we make no particular annotation. Are grouped under the label *holder*, the widest possible explicit mention of the holder: including modifiers, appositions, conjunction, etc., in order to have maximum semantic information.

SOURCE							
Je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8

Figure 12: Holder (SOURCE in french) annotation

HOLDER				
Ich	mag	Pasta	nicht	wirklich
1	2	3	4	5

Figure 13: Holder annotation for german

C'	est	nul	!
1	2	3	4

Figure 14: The holder is the writer of the text: No explicit mention of the holder in the text

Es	ist	nichts	wert	!
1	2	3	4	5

Figure 15: German example with no explicit mention of the holder in the text

SOURCE																		
En	tant	que	cuisinier	amateur	qui	a	de	l´	expérience	,	je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 16: The holder annotation include the widest possible explicit mention

HOLDER							HOLDER				
Als	ein	Amateurkoch	mit	Erfahrung	,	mag	ich	Pasta	nicht	wirklich	.
1	2	3	4	5	6	7	8	9	10	11	12

Figure 17: Example with a german holder annotation including the widest possible explicit mention

3.5.1.2 Target

As for the holder, we annotate as *target*, the widest possible explicit mention of the opinion, sentiment and emotion target. If there is multiple target for the same OSEE expression, we create multiple group target.

CIBLE			CIBLE			CIBLE						
Les	lynx	,	les	loups	,	les	tortues	sont	des	éspéces	protégées	.
1	2	3	4	5	6	7	8	9	10	11	12	13

Figure 18: Annotation with multiple target group

3.5.1.3 Negation

This group refers to markers of negation.

		NEGATION		NEGATION				
Le	serpent	n'	est	pas	une	espèce	protégée	.
1	2	3	4	5	6	7	8	9

Figure 19: Negation group

3.5.1.4 Modifier

This group refers to modifier markers.

						MODIFIEUR							
La	Sardine	l'	un	des	poissons	les	plus	en	danger	en	Méditerranée	http	:
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 20: Negation group

3.5.1.5 Recipient

We group under the *recipient* lablel the explicit mention of the recipient of the OSEE expression. This group will be used in the case the message is intended for someone.

DESTINATAIRE										
Mme	Ségolène	Royale	vous	êtes	priée	de	respecter	les	loups	.
1	2	3	4	5	6	7	8	9	10	11

Figure 21: Recipient group

3.5.1.6 Opinion, Sentiment, Emotion Expression (OSEE)

Annotation with fine-grained affective and semantic classes. The OSEE group consists of the span of text which the semantic value corresponds to the OSEE expression. It will be annotated with one of the 22 semantic or affective categories given in the following table. The definitions of the categories are given afterwards, with annotations examples.

#	Generic label	Dim.	uComp specific semantic category
1	NEGATIVE SURPRISE	e-	negative surprise / negative amazement
2	DISCOMFORT	e-	discomfort / disturbance / embarassment / guilt
3	FEAR	e-	shyness / worry / apprehension / alarm fear / terror
4	BOREDOM	e-	boredom
5	DISPLEASURE	e-	displeasure / deception / abuse
6	SADNESS	e-	sadness / resignation / despair / sorrow / hopelessness
7	ANGER	e-	impatience / annoyance / irritation / nervousness / anger / exasperation
8	CONTEMPT	e-	reluctance / contempts / disdain / blame / disgust / hate
9	DISATISFACTION	s-	disappointment / disatisfaction / discontent / shame
10	DEVALORIZATION	o-	disinterest / devalorization / depreciation
11	DISAGREEMENT	o-	disapproval / disagreement
12	VALORIZATION	o+	interest / valorization / appreciation
13	AGREEMENT	o+	understanding / approval / agreement
14	SATISFACTION	s+	satisfaction / contentment / pride
15	POSITIVE SURPRISE	e+	positive surprise / positive amazement
16	APPEASEMENT	e+	relief / appeasement / peacefullness forgiveness / thankfullness
17	PLEASURE	e+	pleasure / entertainment / enjoyment / joy / happiness / euphoria / play
18	LOVE	e+	love / affection / care / tenderness / fondness / kindness / attachment / devotion / passion / envy / desire
19	INSTRUCTION	i	recommandation / suggestion / instruction / order /command
20	POSITIVE	+	underspecification for an positive OSEE which cannot be ascribed to any of the finer categories above
21	NEGATIVE	-	underspecification for an negative OSEE which cannot be ascribed to any of the finer categories above
22	UNKNOWN	?	some OSE is present but it is either missing from the list above or difficult to determine because it is composed of a mix of many different sentiments, emotions or opinions.

Table 31: uComp fine grained semantic categories of OSEE.
e=emotion, s=sentiment, o=opinion, i=information, +=positive valence, and -=negative valence

DISPLEASURE

Definition : Negative emotion resulting from the occurence of an unwanted event.
Annotation example:

															DÉPLAISIR				
Un	lion	en	cage	,	un	singe	et	deux	serpents	:	une	soirée	cirque	qui	passe	mal	au	Stamp	#Waterloo
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figure 22: DISPLEASURE group

DISCOMFORT

Definition: Negative emotion resulting from the occurence of an unwanted event, which succite an intention to action in order to remedy to the event.
Annotation example:

					DÉRANGEMENT
Les	éoliennes	vraiment	source	de	nuisances
1	2	3	4	5	6

Figure 23: DISCOMFORT group

CONTEMPT

Definition : Negative emotion resulting from our knowledge about an entity (a person, an organisation), that is in opposition to our desires.
Annotation example:

	MÉPRIS			MÉPRIS							
Les	losers	et	les	saloppes	,	deux	espèces	en	voie	d'	extinction
1	2	3	4	5	6	7	8	9	10	11	12

Figure 24: CONTEMPT group

NEGATIVE SURPRISE

Definition : Negative emotion resulting from the occurence of an unwanted event and unexpected event.
Annotation example:

SURPRISE NÉGATIVE																	
Mauvaise	nouvelle	:	le	ministre	Henry	a	accordé	le	permis	unique	à	Spe	Luminus	pour	pour	5	éoliennes
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Figure 25: NEGATIVE SURPRISE group

FEAR

Definition : Negative emotion resulting from the realization or the eventual realization of an unwanted event.
Annotation example:

								PEUR			
La	sardine	l'	un	des	poissons	les	plus	en	danger	en	Méditerranée
1	2	3	4	5	6	7	8	9	10	11	12

Figure 26: FEAR group

ANGER

Definition: Negative emotion resulting from the realization of an not desirable event, which may raise or not a reaction.
Annotation example:

		COLÈRE								COLÈRE	
Vent	de	colère	sur	nos	villages	.	Éoliennes	:	l'	arnaque	totale
1	2	3	4	5	6	7	8	9	10	11	12

Figure 27: ANGER group

BOREDOM

Definition : Negative emotion resulting from the knowledge that there is no desirable event for the experiencer.

SADNESS

Definition : Negative emotion resulting from the non realization of a desirable event whose realization is possible or impossible in the future.
Annotation example:

																			TRISTESSE	
Je	vois	des	trucs	clignoter	mais	je	sais	pas	si	c'	est	des	éoliennes	ou	des	feux	d'	artifices	c	trist
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Figure 28: SADNESS group

PLEASURE

Definition: positive emotion resulting from the realization of desirable event.
Annotation example:

	PLAISIR					
J'	aime	trop	voir	les	éoliennes	!
1	2	3	4	5	6	7

Figure 29: PLEASURE group

APPEASEMENT

Definition : positive emotion resulting from the realization of an intention following the realization of not desirable event.
Example:

LOVE

Definition : positive emotion prompted by the desire of a person or an animal.
Annotation example:

	AMOUR			AMOUR							
L'	amour	et	la	fidelité	sont	des	espéces	en	voie	de	disparition
1	2	3	4	5	6	7	8	9	10	11	12

Figure 30: LOVE group

POSITIVE SURPRISE

Definition: positive emotion resulting from the realization od a desirable and unexpected event.
Annotation example:

SURPRISE POSITIVE												
Bonne	nouvelle	pour	Brest	qui	construira	les	jackets	!	Saint	-	Brieuc	.
1	2	3	4	5	6	7	8	9	10	11	12	13

Figure 31: POSITIVE SURPRISE group

SATISFACTION

Definition: positive sentiment resulting from the realization of an intention resulting from a desire.
Example:

DISATISFACTION

Definition: negative sentiment resulting from the non-realization of an intention resulting from a desire.
Example:

AGREEMENT

Definition: positive opinion, the person is agree with at least another person.
Example:

VALORIZATION

Definition : positive opinion, the person desire an entity (person, organization, service, object, event, etc.) and has the intention to do an action in favor of this entity.
Annotation example:

																VALORISATION	
#tortueluth	espèce	menacée	"	Vu	du	ciel	"	#Gabon	,	les	héros	de	la	nature	"	super	doc
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Figure 32: VALORIZATION group

DISAGREEMENT

Definition: negative opinion, the person is not agree.
Annotation example:

DÉSACCORD													
Non	aux	éoliennes	de	la	ferme	du	Torpt	à	Tourville	et	St	-	Meslin
1	2	3	4	5	6	7	8	9	10	11	12	13	14

Figure 33: DISAGREEMENT group

DEVALORIZATION

Definition: negative opinions, the person not desire an entity (person, organization, service, object, event, etc.) and don't has the intention to do an action in favor of this entity.
Annotation example:

	DÉVALORISATION						DÉVALORISATION		DÉVALORISATION		DÉVALORISATION							DÉVALORISATION	
Les	saloperies	promues	par	Royal	:	c'	bruyant	,	laid	,	dévalorisant	pour	le	foncier	,	et	même	pas	efficace
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Figure 34: DEVALORIZATION group

INSTRUCTION

Definition: instruction, order, recommandation, etc.
Annotation example:

INSTRUCTION							
Pensez	à	recycler	vos	,	bouteilles	vides	.
1	2	3	4	5	6	7	8

Figure 35: INSTRUCTION group

NEGATIVE

Definition: negative opinion, sentiment or emotion
Annotators can use this group, when the context is insufficient to identify correctly the exact affective and semantic class od the OSEE.
Annotation example:

	NÉGATIF					NÉGATIF					
Ivoire	Maudit	Kenya	:	le	pire	massacre	de	rhinocéros	depuis	1988	#Quebec
1	2	3	4	5	6	7	8	9	10	11	12

Figure 36: NEGATIVE group

POSITIVE

Definition: positive opinion, sentiment or emotion
Annotators can use this group, when the context is insufficient to identify correctly the exact affective and semantic class of the OSEE.

UNKNOWN

Definition: undetermined opinion, sentiment or emotion, maybe resulting of too many opinions, sentiments, or emotions expressed together. Annotators can use this group, when the context is insufficient to identify correctly the exact affective and semantic class of the OSEE.

3.5.1.7 Global OSE

All, messages to be annotated start by the token *Global_ OSE*. The annotator must associate to this token the global semantic category of the message.
Annotation example:

VALORISATION				DÉVALORISATION										VALORISATION
OSE_Globale	Même	si	je	déteste	les	éoliennes	,	il	faut	avouer	qu'	elles	sont	utiles
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Figure 37: Global OSE of the message: VALORIZATION

Annotation example:

DÉVALORISATION			VALORISATION				
OSE_Globale	Je	n'	aime	pas	les	éoliennes	.
1	2	3	4	5	6	7	8

Figure 38: Global OSE of the message: DEVALORIZATION

3.5.2 Relations

3.5.2.1 SAY

The relation *SAY* connects the group *holder* to the group *OSSE*.

DIT							
SOURCE	OSEE						
Je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8

Figure 39: The SAY relation

3.5.2.2 ABOUT

The relation *ABOUT* connects the group *OSSE* to group *Target*

	SUR						
		VALORISATION			CIBLE		
Je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8

Figure 40: The ABOUT relation

SUR						
			SUR			
CIBLE			CIBLE		VALORISATION	
le	loup	est	une	espéce	protégée	.
1	2	3	4	5	6	7

Figure 41: The ABOUT relation

3.5.2.3 MOD

The relation *MOD* connects the group *OSSE* to the group *MODIFIER*

		MOD					
		VALORISATION		MODIFIEUR			
Je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8

Figure 42: The MOD relation

	MOD				
	VALORISATION	MODIFIEUR			
J'	aime	vraiment	les	pâtes	.
1	2	3	4	5	6

Figure 43: The MOD relation

3.5.2.4 NEG

The relation *NEG* connects the group *OSSE* to the group *NEG*

	NEG						
		NEG					
	NÉGATION	VALORISATION	NÉGATION				
Je	n'	aime	pas	vraiment	les	pâtes	.
1	2	3	4	5	6	7	8

Figure 44: The NEG relation

	NEG					
	VALORISATION		NÉGATION			
J'	aime	vraiment	pas	les	pâtes	.
1	2	3	4	5	6	7

Figure 45: The NEG relation

3.5.2.5 RECEPTOR

The relation *RECEPTOR* connects the group *OSSE* to the group *RECIPIENT*

RECEPTEUR										
DESTINATAIRE			INSTRUCTION							
Mme	Ségolène	Royale	vous	êtes	priée	de	respecter	les	loups	.
1	2	3	4	5	6	7	8	9	10	11

Figure 46: The RECEPTOR relation

3.6 Expert Annotation Software

A specific annotation stand-alone annotation interface with the accompanying text importer has been developped by customizing the Pasta annotation interface for syntactic annotation developped in the PASSAGE project ([Vilnat et al., 2010]). The text importer is written in C++ and converts raw text into xml suitable to be imported in Pasta. The Pasta interface is written in Java.

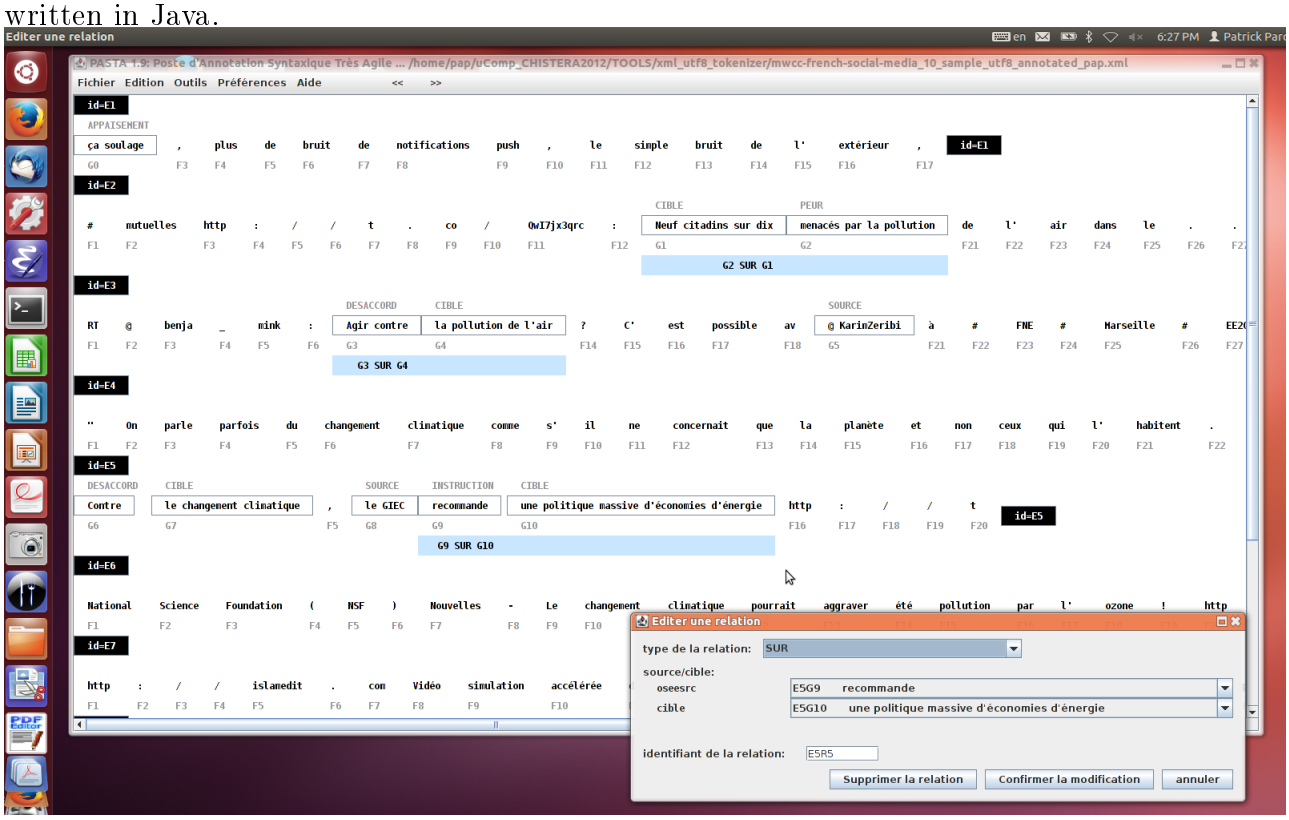

Figure 47: Pasta-uComp annotation interface.

3.7 Campaign Deployment

Contact have been taken with the organization committee of the DEFT series of evaluation campaign for text mining (see `http://deft.limsi.fr/2014/`), the uComp evaluation campaign will be deployed in the 2015 DEFT issue, as 3 different tracks of opinion mining in a multilingual set-up with different level of granularity annotation (sentence, chunk and relation level).

References

[Baker et al., 2010] Baker, K., Bloodgood, M., Dorr, B. J., Filardo, N. W., Levin, L. S., and Piatko, C. D. (2010). A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Bessho et al., 2012] Bessho, F., Harada, T., and Kuniyoshi, Y. (2012). Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 227–231, Seoul, South Korea. Association for Computational Linguistics.

[Bird and Liberman, 2000] Bird, S. and Liberman, M. (2000). A Formal Framework for Linguistic Annotation. *Speech Communication*, 33:23–60.

[Boyd-Graber et al., 2012] Boyd-Graber, J., Satinoff, B., He, H., and Daume III, H. (2012). Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea. Association for Computational Linguistics.

[Chiao et al., 2006] Chiao, Y.-C., Kraif, O., Laurent, D., Minh, T., Nguyen, H., Semmar, N., Stuck, F., Véronis, J., and Zaghouani, W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of the Fifth conference on International Language Resources and Evaluation (LREC'06)*, pages 1975–1979, Genoa, Italy.

[Cleverdon, 1960] Cleverdon, C. (1960). The ASLIB Cranfield research project on the comparative efficiency of indexing systems. *ASLIB Proceedings*, 12:421–431. ISSN: 0001-253X / DOI: 10.1108/eb049778.

[Edmonds and Kilgarriff, 2002] Edmonds, P. and Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Nat. Lang. Eng.*, 8:279–291.

[Finin et al., 2010] Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles. Association for Computational Linguistics.

[Fort et al., 2014] Fort, K., Adda, G., Sagot, B., Mariani, J., and Couillault, A. (2014). Crowdsourcing for Language Resource Development: Criticisms About Amazon Mechanical Turk Overpowering Use. In Vetulani, Zygmunt and Mariani, Joseph, editor, *Human Language Technology Challenges for Computer Science and Linguistics*, pages 303–314. Springer International Publishing.

[Fraisse and Paroubek, 2013] Fraisse, A. and Paroubek, P. (2013). uComp Deliverable D5.1 - Requirements of Affective Knowledge Extraction. Technical report, LIMSI-CNRS.

[Gindl et al., 2010] Gindl, S., Weichselbraun, A., , and Scharl, A. (2010). Cross-domain contextualization of sentiment lexicons. In *Proceedings of the Eur. Conf. on Artif. Intelligence (ECAI-2010)*, pages 771–776. IOS Press.

[Grady and Lease, 2010] Grady, C. and Lease, M. (2010). Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 172–179, Los Angeles. Association for Computational Linguistics.

[Higgins et al., 2010] Higgins, C., McGrath, E., and Moretto, L. (2010). Mturk crowdsourcing: A viable method for rapid discovery of arabic nicknames? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 89–92, Los Angeles. Association for Computational Linguistics.

[Hsueh et al., 2009] Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado. Association for Computational Linguistics.

[Hu et al., 2011] Hu, C., Resnik, P., Kronrod, Y., Eidelman, V., Buzek, O., and Bederson, B. B. (2011). The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using haitian creole emergency sms messages. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 399–404, Edinburgh, Scotland. Association for Computational Linguistics.

[Koller et al., 2010] Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2010). Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In *Proceedings of the International Natural Language Generation Conference (INLG)*, Dublin.

[Lafourcade and Fort, 2014] Lafourcade, M. and Fort, K. (2014). Propa-l: a semantic filtering service from a lexical network created using games with a purpose. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Larson, 2013] Larson, M. (2013). Responsible crowdsourcing. `http://homepage.tudelft.nl/q22t4/lib/EthicsOfCrowdsourcing_v2.pdf`. Delft University of Technology, seminar on "Crowdsourcing: From Theory to Practice and Long-Term Perspectives", Dagstuhl, September 1-4, 2013.

[Lease, 2013a] Lease, M. (2013a). Crowdsourcing & ethics: a few thoughts and refences. `http://fr.slideshare.net/mattlease/crowdsourcing-ethics-a-few`. Extracts and addendums from an earlier talk, for those interested in ethics and related issues in regard to crowdsourcing, particularly research uses. Slides updated Sept. 2, 2013.

[Lease, 2013b] Lease, M. (2013b). Crowdsourcing for information retrieval: From statistics to ethics presentation transcript. `http://fr.slideshare.net/mattlease/lease-statisticsethics`. Revised October 27, 2013. Talk at UC Berkeley (October 21, 2013), Syracuse University (October 28, 2013).

[Madnani et al., 2011] Madnani, N., Chodorow, M., Tetreault, J., and Rozovskaya, A. (2011). They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 508–513, Portland, Oregon, USA. Association for Computational Linguistics.

[Magnini et al., 2008] Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., Mazzei, A., Lombardo, V., Bertagna, F., Calzolari, N., Toral, A., Lenzi, V. B., Sprugnoli, R., and Speranza, M. (2008). Evaluation of Natural Language Tools for Italian: EVALITA 2007. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Manning and Schütze, 2002] Manning, C. D. and Schütze, H. (2002). *Foundation of Statistical Natural Language Processing*. Massachusetts institute of Technology Press, 5th edition.

[Marcus et al., 1993] Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.

[de la Clergerie et al., 2008] de la Clergerie, E., Hamon, O., Mostefa, D., Ayache, C., Paroubek, P., and Vilnat, A. (2008). PASSAGE: from French Parser Evaluation to Large Sized Treebank. In ELRA, editor, *In proceedings of the sixth international conference on Language Resources and Evaluation (LREC)*, Marrakech, Morroco.

[Munro et al., 2010] Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, Los Angeles. Association for Computational Linguistics.

[Negri et al., 2011] Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., and Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[Pak et al., 2014] Pak, A., Paroubek, P., Fraisse, A., and Francopoulo, G. (2014). *Human Language Technology Challenges for Computer Science and Linguistics*, chapter Normalization of Term Weighting Scheme for Sentiment Analysis. Springer International Publishing Switzerland. *In print*.

[Paroubek, 2007] Paroubek, P. (2007). *Evaluation of Text and Speech Systems*, volume 36 of *Text, Speech and Language Technology*, chapter Evaluating Part Of Speech Tagging and Parsing, pages 97–116. Kluwer Academic Publisher. ISBN-10: 1-4020-5815-2, ISBN-13: 978-1-4020-5815-8.

[Paroubek et al., 2007] Paroubek, P., Chaudiron, S., and Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *Traitement Automatique des Langues (TAL)*, 48(1):7–31.

[Paroubek et al., 1998] Paroubek, P., Lecomte, J., Adda, G., Mariani, J., and Rajman, M. (1998). The GRACE French Part-Of-Speech Tagging Evaluation Task. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 433–441, Granada, Spain. ELDA.

[Paroubek et al., 2009] Paroubek, P., de la Clergerie, E., Loiseau, S., Vilnat, A., and Francopoulo, G. (2009). The PASSAGE Syntactic Representation. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, pages 91–102, Gröningen. Netherlands Graduate Schools of Linguistics (LOT).

[Post et al., 2012] Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

[Prabhakaran et al., 2012] Prabhakaran, V., Bloodgood, M., Diab, M., Dorr, B., Levin, L., Piatko, C. D., Rambow, O., and Van Durme, B. (2012). Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea. Association for Computational Linguistics.

[Roth and Sammons, 2008] Roth, D. and Sammons, M. (2008). A Unified Representation and Inference Paradigm for Natural Language Processing. Technical Report UIUCDCS-R-2008-2969, UIUC Computer Science Department.

[Rumshisky, 2011] Rumshisky, A. (2011). Crowdsourcing word sense definition. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 74–81, Portland, Oregon, USA. Association for Computational Linguistics.

[Rumshisky et al., 2012] Rumshisky, A., Botchan, N., Kushkuley, S., and Pustejovsky, J. (2012). Word sense inventories by non-experts. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors,

Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).

[Rumshisky et al., 2009] Rumshisky, A., Moszkowicz, J., , and Verhagen, M. (2009). The holy grail of sense definition: Creating a sense-disambiguated corpus from scratch. In *In Proceedings of 5th International Conference on Generative Approaches to the Lexicon*, Pisa, Italy.

[Safire, 2009] Safire, W. (2009). On language. *New York Times Magazine*.

[Sagot, 2010] Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

[Sayeed et al., 2011] Sayeed, A., Rusk, B., Petrov, M., Nguyen, H., Meyer, T., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 69–77, Portland, OR, USA. Association for Computational Linguistics.

[Sayeed et al., 2010] Sayeed, A. B., Meyer, T. J., Nguyen, H. C., Buzek, O., and Weinberg, A. (2010). Crowdsourcing the evaluation of a domain-adapted named entity recognition system. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 345–348, Los Angeles, California. Association for Computational Linguistics.

[Scharl, 2014] Scharl, A. (2014). Domain-specific content repository. Technical report, MODUL University.

[Schmitz and Le Roux, 2008] Schmitz, S. and Le Roux, J. (2008). Feature Unification in TAG Derivation Trees. In Gardent, C. and Sarkar, A., editors, *TAG+9*, pages pages 141–148, Tübingen, Allemagne. 12 pages, 4 figures.

[Unnikrishnan et al., 2007] Unnikrishnan, R., Pantofaru, C., and Hebert, M. (2007). Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944.

[Verhagen et al.,] Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. SemEval-2007 - 15: TempEval Temporal Relation Identification.

[Vertanen and Kristensson, 2011] Vertanen, K. and Kristensson, P. O. (2011). The imagination of crowds: Conversational aac language modeling using crowdsourcing and large data sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 700–711, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[Vilain et al., 1995] Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52, Columbia, Maryland, USA. ACL.

[Vilnat et al., 2010] Vilnat, A., Paroubek, P., de la Clergerie, E. V., Francopoulo, G., and Guénot, M.-L. (2010). PASSAGE Syntactic Representation: a Minimal Common Ground for Evaluation. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Watson et al., 2005] Watson, R., Carroll, J., and Briscoe, T. (2005). Efficient extraction of grammatical relations. In *Proceedings of the Ninth International Workshop on Parsing Technologies (IWPT)*, pages 160–170, Vancouver. Association for Computational Linguistics.

[Zaidan and Callison-Burch, 2011] Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

[Zeichner et al., 2012] Zeichner, N., Berant, J., and Dagan, I. (2012). Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–160, Jeju Island, Korea. Association for Computational Linguistics.