DELIVERABLE SUBMISSION SHEET

To: Susan Fraser *(Project Officer)*

EUROPEAN COMMISSION
Directorate-General Information Society and Media
EUFO 1165A
L-2920 Luxembourg

From:

Project acronym:	PHEME	Project number:	611233
Project manager:	Kalina Bontcheva		
Project coordinator	The University of Sheffield (USFD)		

The following deliverable:

Deliverable title:	Linguistics Pre-processing Tools and Ontological Models of Rumours and Phemes
Deliverable number:	D2.2
Deliverable date:	31 December 2014
Partners responsible:	USAAR
Status:	☒ Public ☐ Restricted ☐ Confidential

is now complete. ☒ It is available for your inspection.
☒ Relevant descriptive documents are attached.

The deliverable is:

☐ a document
☐ a Website (URL: .....)
☐ software (.....)
☐ an event
☒ other (.....Prototype.....)

Sent to Project Officer:	Sent to functional mail box:	On date:
Susan.Fraser@ec.europa.eu	*CNECT-ICT*-611233@ec.europa.eu	08 January 2015

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)

Computing Veracity Across Media, Languages, and Social Networks

D2.2 Linguistic Pre-processing Tools and Ontological Models of Rumours and Phemes

Thierry Declerck (Universität des Saarlandes)
Petya Osenova (Ontotext AD)
Leon Derczynski (University of Sheffield)

Abstract

FP7-ICT Strategic Targeted Research Project PHEME (No. 611233)
Deliverable D2.2 (WP 2)

In this deliverable we report on work done in the context of task 2.2 "Ontological modelling" and task 2.3 "Multilingual Pre-processing" of WP2 "Ontologies, Multilinguality, and Spatio-Temporal Grounding" of the Pheme project. The aim of those taks was 1) to build ontological models of veracity, misinformation, social and information diffusion networks, rumours, disputed claims, temporal validity of statements, and user online behavior; and 2) to compare adopt and adapt the necessary linguistic pre-processing tools for Bulgarian, English and German, including language identification, POS tagging, chunking, dependency parsing, entity and relation recognition, and LOD-based entity disambiguation. These adapted tools will be used to generate linguistic and semantic features for the methods to be deployed in WP3 "Contextual Interpretation" and WP4 "Detecting Rumours and Veracity".

Keyword list: ontologies, text pre-processing, social media, rumours

---

Nature: Prototype
Dissemination: PU
Contractual date of delivery: 31 Dec 2014
Actual date of delivery: 08 Jan 2015
Reviewed By: Kalina Bontcheva
Web links: http://www.pheme.eu/software-downloads/

PHEME Consortium

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP, UK
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Universitaet des Saarlandes
Language Technology Lab
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Executive Summary

In this deliverable we report on work done in the context of task 2.2 “Ontological modelling” and task 2.3 “Multilingual Pre-processing” of WP2 “Ontologies, Multilinguality, and Spatio-Temporal Grounding” of the Pheme project. The aim of those taks was 1) to build ontological models of veracity, misinformation, social and information diffusion networks, rumours, disputed claims, temporal validity of statements, and user online behavior; and 2) to compare adopt and adapt the necessary linguistic pre-processing tools for Bulgarian, English and German, including language identification, POS tagging, chunking, dependency parsing, entity and relation recognition, and LOD-based entity disambiguation.

These adapted tools will be used to generate linguistic and semantic features for the methods to be deployed in WP3 “Contextual Interpretation“ and WP4 “Detecting Rumours and Veracity“.

Contents

1 Introduction

In this deliverable we report on work done in the context of task 2.2 "Ontological modelling" and task 2.3 "Multilingual Pre-processing" of WP2 "Ontologies, Multilinguality, and Spatio-Temporal Grounding" of the Pheme project. The aim of those taks was 1) to build ontological models of veracity, misinformation, social and information diffusion networks, rumours, disputed claims, temporal validity of statements, and user online behavior; and 2) to compare adopt and adapt the necessary linguistic pre-processing tools for Bulgarian, English and German, including language identification, POS tagging, chunking, dependency parsing, entity and relation recognition, and LOD-based entity disambiguation.

In this document we report mainly on the general aspects of ontology modelling for Pheme, and this is described in details in section 2.

In section 3 we describe work done on the evaluation existing tools for processing text in the three languages of the project, Bulgarian, English and German, in order to see how they can perform on user generated content. We consider here various levels of text processing, from language independent language identification up to part-of-speech tagging and dependency analysis, also considering Named Entity recognition and linking.

Relevance to project objectives

The work on modeling and on adapting text analysis tools is embedded in some general technological objectives of Pheme, which we list here:

Develop innovative, multilingual methods for cross-media detection of phemes

Task 2.2 and 2.3 are supporting the goal of extracting and reasoning about multiple truths (e.g. in controversies) and taking into account *the context in which phemes originated and spread* (e.g. the trustworthiness and influence of their sources, etc.).

In order to support the multilingual analysis of various types of documents published in the social media, work in task 2.3 has investigated how current natural language processing tools have to be adapted to user generated content and proposes a first evaluation of such an adaptation, so that the project knows what quality can be expected from such adapted tools.

Integrate large-scale, a priori knowledge from Linked Open Data (LOD)

This objective aims at improving pheme identification methods in specific application domains. The first version of the Pheme ontology described in this deliverable is mediating with generic domain and use case specific ontological model available in the Linked Data.

Model pheme spread dynamics over time, and within and across social networks and media.

The Pheme ontology establishes models of rumour and information spread and so it helps cross-referencing and comparing facts and rumours across different types of publications, including user generated content in the social media framework.

Relation to other workpackages

The developed ontologies and the adapted text processing tools will be used to generate linguistic and semantic features for the methods to be deployed in WP3 "Contextual Interpretation" and WP4 "Detecting Rumours and Veracity".

Updates of the ontology modelling will be provided to follow new requirements by the two use cases of Pheme defined in WP7 "Veracity Intelligence in Patient Care" and WP8 "Digital Journalism Use Case", and will be reported in the corresponding deliverables.

2 Knowledge Modeling in Pheme

Ontologies are nowadays widely used as conceptualization models of domains of applications, both in the areas of knowledge representation and Natural Language Processing (NLP). Ontologies act as controlling mechanisms over the relevant data and as means for ensuring adequate inference mechanisms over the facts. For that reason, Pheme relies on the usage of focused ontologies that are modelling the domains of the use cases of the project, but also the types of language data and linguistic phenomena the project is dealing with.

The motivation behind the task T2.2 "Ontological modelling" was to build new and to extend existing ontologies to model veracity (including the temporal validity of statements), misinformation, rumours, and disputed claims. The ontologies need also to model social and information diffusion networks, users (content authors, receivers and diffusers), lexicalizations and sentiment entities, events and relations. A goal here is also to be able to map and compare extracted statements to data sets published in the Linked Open Data (LOD) framework, and more specifically to authoritative sources in the LOD.

2.1 The Pheme ontology

We decided to develop a top Pheme ontology that reflects closely the Annotation Scheme for Social Media Rumours, presented in the Pheme deliverable D2.1 "Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages". The present ontology version reflects current version of this scheme. It must be noted, however, that the ontology can be easily adjusted towards the incorporation of any further developments of the scheme.

2.1.1 The basis: PROTON

PROTON[1] was selected as the top ontology for the following reasons: it is an in-house resource of the project partner ONTO, and it supports the linking to DBpedia[2] and other LOD datasets (FreeBase[3], Geonames[4], etc.). The PROTON (PROTo ONtology) ontology has been developed in the past SEKT project[5] as a light-weight upper-level ontology, serving as a modelling basis across different tasks and domains.

PROTON is applied to integrate and access multiple datasets is the FactForge.net[6] – a public service, which allows the user or the application to explore and query efficiently a dataset of 3 billion statements, combining 10 of the most central LOD datasets. The user can either use the original vocabularies of the data sets or the

[1] http://www.ontotext.com/proton-ontology/
[2] http://dbpedia.org/About
[3] https://www.freebase.com/
[4] http://www.geonames.org/
[5] http://www.ontotext.com/research/sekt
[6] http://www.ontotext.com/factforge-links/

PROTON primitives. In the latter case the user does not have to deal with the peculiarities of the different datasets. FactForge users benefit also from reasoning on the basis of the PROTON semantics and the owl:sameAs statements made between the datasets.

The PROTON ontology contains more than 500 classes and 200 properties, providing coverage of the general concepts necessary for a wide range of tasks, including semantic annotation, indexing, and retrieval. The design principles can be summarized as follows:

- domain-independence;
- light-weight logical definitions;
- alignment with popular metadata standards;
- good coverage of named entity types and concrete domains (i.e. modelling of concepts such as people, organizations, locations, numbers, dates, addresses); and
- good coverage of instance data in Linked Open Data Reason-able view Fact Forge.

The ontology is encoded in OWL 2 RL[7] sub-set (that is also eligible OWL 2DL) and split into two modules: *Top* and *Extension*. A snapshot of the top of PROTON class hierarchy and some of the extension is given in Figure 1.

[7] http://www.w3.org/TR/owl2-profiles/#OWL_2_RL

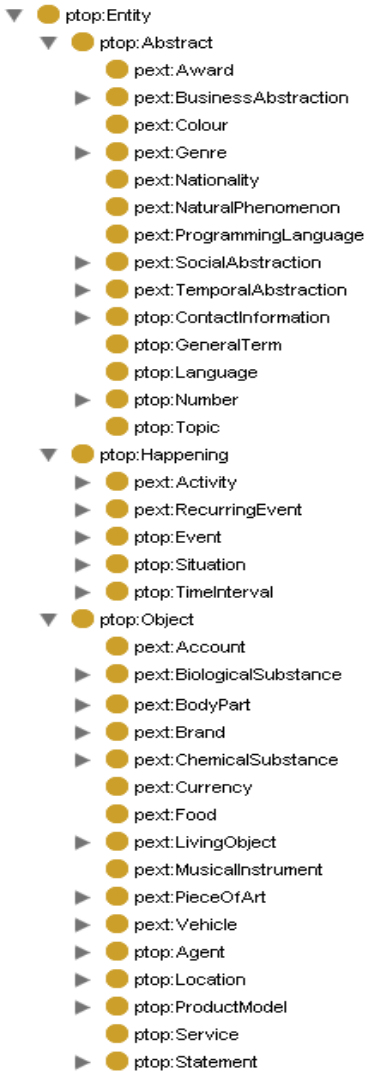

Figure 1: A view of top and some extended part of the PROTON class hierarchy

The top part starts with the prefix *ptop*, which means 'the top of PROTON'. For example, the top class *InformationResource* looks like this:

```
ptop:InformationResource
      rdf:type owl:Class ;
      rdfs:comment "InformationResource denotes an information
resource with identity, as defined in Dublin Core (DC2003ISO).
InformationResource is considered any communication or message that
is delivered or produced, taking into account the specific intention
of its originator, and also the supposition (and anticipation) for a
particular audience or counter-agent in the process of communication
(i.e. passive or active feed-back)."@en ;
      rdfs:label "Information Resource"@en ;
      rdfs:subClassOf ptop:Statement .
```

The extensions have been made to handle the Linked Open Data categories. They start with the prefix *pext*, which means 'the extension of PROTON'. For example, the extended concept of *Artery* looks like this:

```
pext:Artery
      rdf:type owl:Class ;
      rdfs:comment "Any artery as a part of the body."@en ;
```

```
rdfs:label "Artery"@en ;
rdfs:subClassOf pext:BodyPart .
```

PROTON is a suitable ontology for tasks related to automatic entity recognition and more generally Information Extraction (IE) from text, for the sake of semantic annotation (metadata generation). It also provides solid basis of data integration and RDF-based extraction, transformation and loading (ETL) of data.

Apart from that, other existing related ontologies were considered: The LivePost Ontology (Scerri et al, 2012), which models interlinked authoritative sources; User Behaviour Ontology (Angeletou et al, 2011); LOD sets, such as DBpedia, GeoNames, etc.

Through its mapping to PROTON, the proposed Pheme ontology also has the advantage to be connected to FactForge, and thus - to the LOD datasets including: DBPedia, FreeBase and GeoNames.

FactForge incorporates the following elements:

- LOD datasets loaded in one RDF repository
- LOD ontologies
- A unification ontology (PROTON) to cover the conceptualizations within the different LOD datasets.

2.1.2 Pheme extensions

Since there is an overlapping of the conceptualizations in the different ontologies, the knowledge in the project was divided on two levels – *common world knowledge* (including different datasets from LOD) and *Pheme knowledge*, extracted from a set of tweets, describing knowledge about a rumour and its development in time.

The Pheme ontology followed the division of tweets in *Source* and *Response*. They have different conceptual models with some overlapping features. They share for example the sub-hierarchy "**Support"**. Below, in Figure 2 and Figure 3 we display the scheme models for "Source" and "Response" Tweets as well as their shared **Support** element.

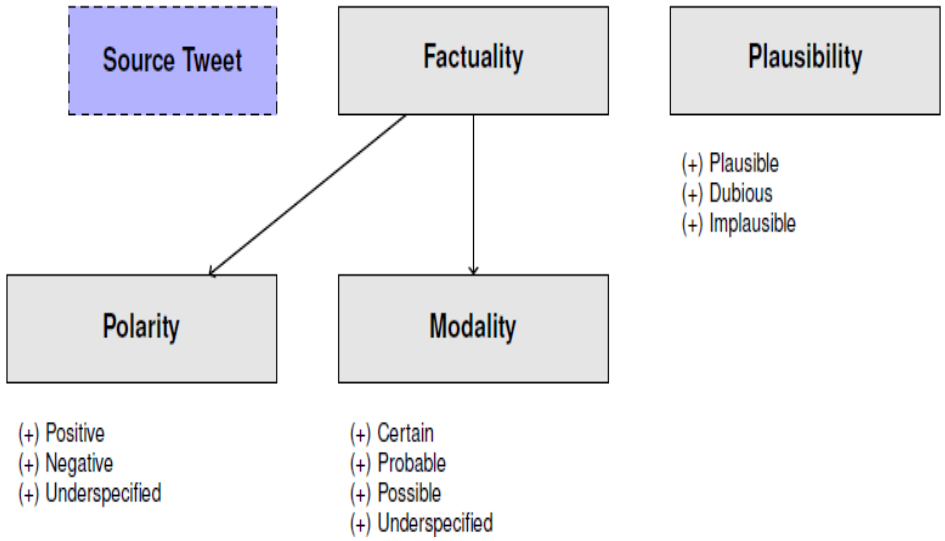

Figure 2: Scheme model for "Source Tweet" in the Pheme ontology

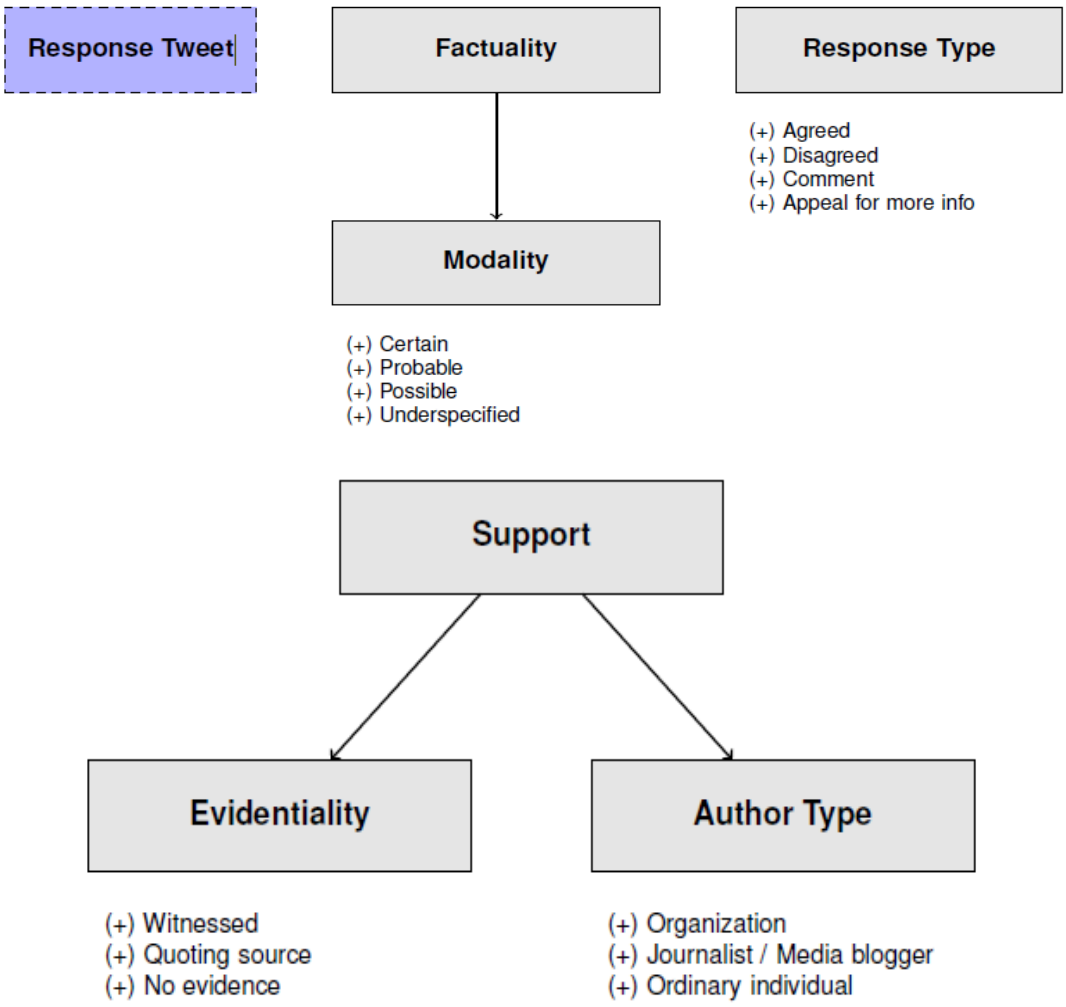

Figure 3: Scheme model for "Response Tweet " in the Pheme ontology

Figure 4 below shows the Pheme classes. We consider a **Pheme** as a **Statement** which is expressed in the texts. As a statement, a **Pheme** has its *lifecycle*, *topic*, and *truth* evaluation. The *lifecycle* defines its author (**Agent**), means of creation (**Statement**), time span (**datetime**). *Topic* is defined as a set of RDF statements. *Truth* evaluation is defined by truthfulness and deception assigned to the topic.

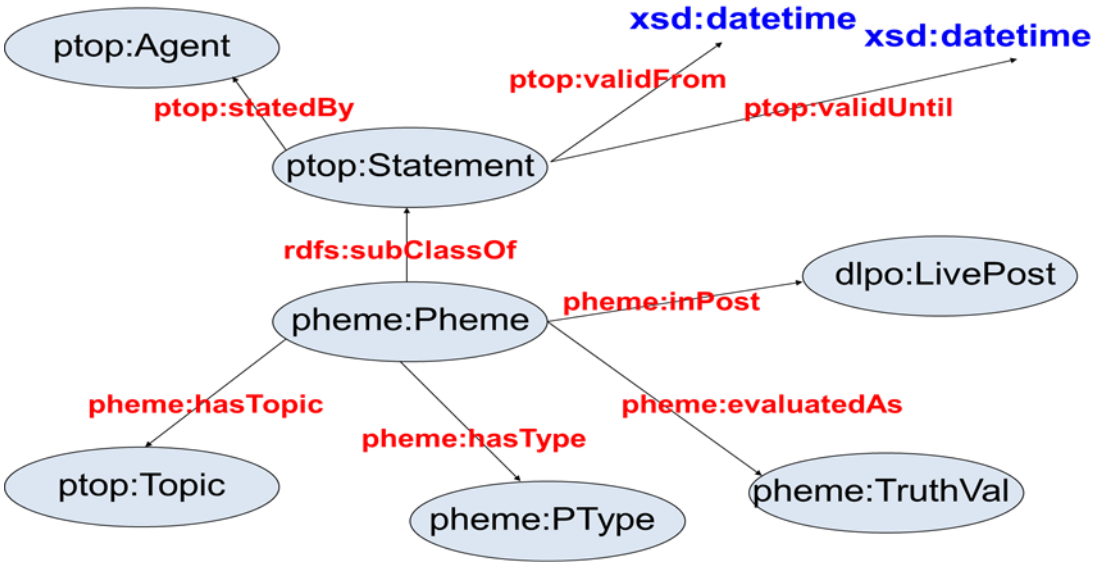

Figure 4: The Classes of the Pheme Ontology

Some explanations to Figure 4: The classes and properties, marked with the prefix *ptop*, come from the PROTON top ontology (classes: **Topic, Statement, Agent**; properties: **statedBy, validFrom, validUntil**). The classes and properties, marked as *Pheme*, are defined in the Pheme ontology (classes: Pheme, Ptype (=Pheme type), TruthVal; properties: hasTopic, hasType, evaluatedAs, inPost, etc.). The prefix dlpo mean that the LivePost ontology is imported into the Pheme one. In this case this is the LivePost ontology. The prefix *rdfs* means that some class is a subclass of another one, following the rdf schema. The prefix xsd means that some XML schema has been used for typing literal values of datatype properties.

Figure 5, below, displays subclasses of the type **Pheme**. For the moment the unique direct subclass of **Pheme** is **Rumour**. **Rumour** has 4 subclasses: **Speculation**, **Controversy**, **Misinfor(mation)** and **Disinfor(mation)**. Here all the presented classes are specific to the Pheme ontology. There is only one relation defined: subClassOf.

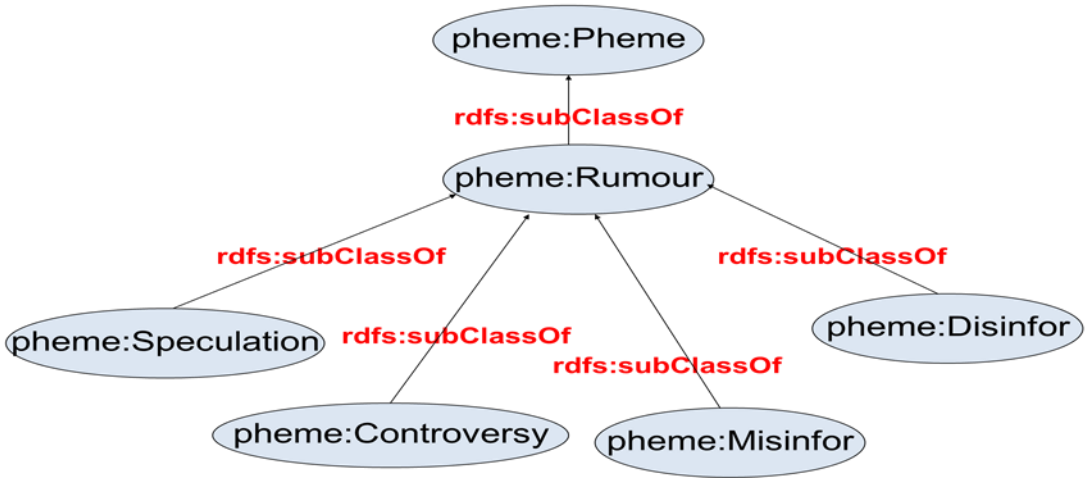

Figure 5: The subclasses of "Pheme"

In Figure 6 below, the Pheme **Tweet class** is introduced as subclass of **InformationResource**. The Tweet class has 2 subclasses: **SourceTweet** and **ResponseTweet**.

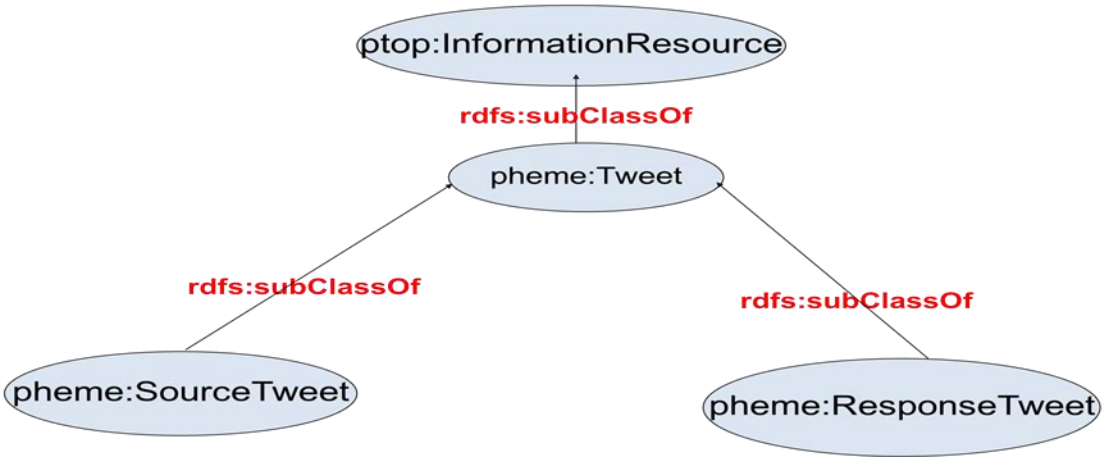

Figure 6: The relation between ptop:InformationResource and the Source and Response Tweet models of Pheme

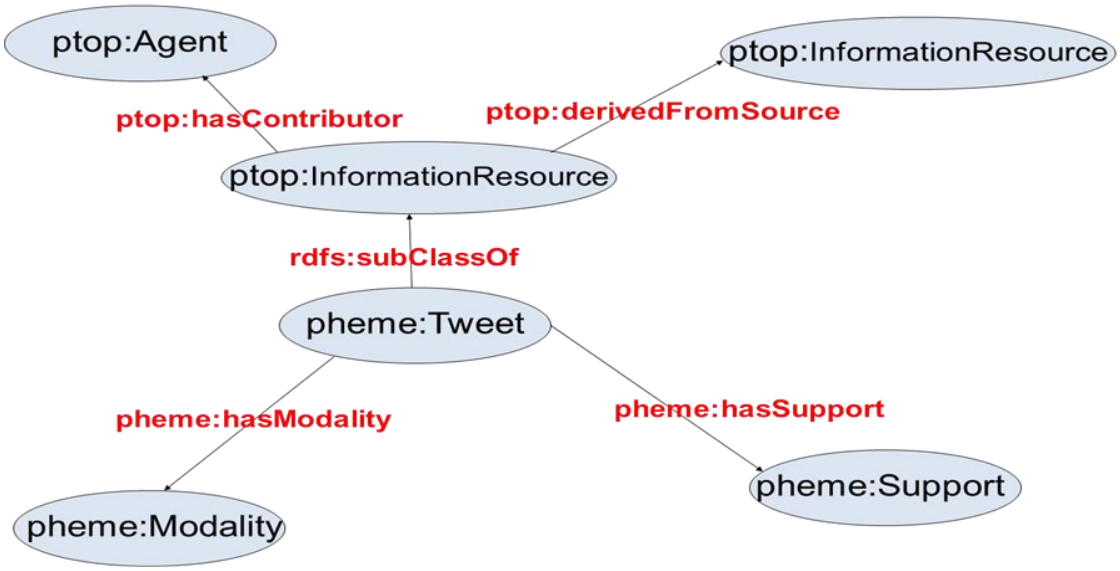

Figure 7: Properties associated with the Tweet class of the Pheme ontology

Figure 7 shows the 5 properties of the Tweet class. Two of them come from the upper ontology PROTON - *hasContributor* (Agent) and *derivedFromSource* (InformationResource); two are specific for Pheme ontology - *hasModality* (Tweet) and *hasSupport* (Tweet). The fifth one defines the relation *subClassOf* with the prefix *rdfs – Tweet is a subClassOf InformationResource*.

Figure 8 displays the four properties of the SourceTweet class: *hasModality*, *hasSupport*, *hasPolarity*, *hasPlausability*.

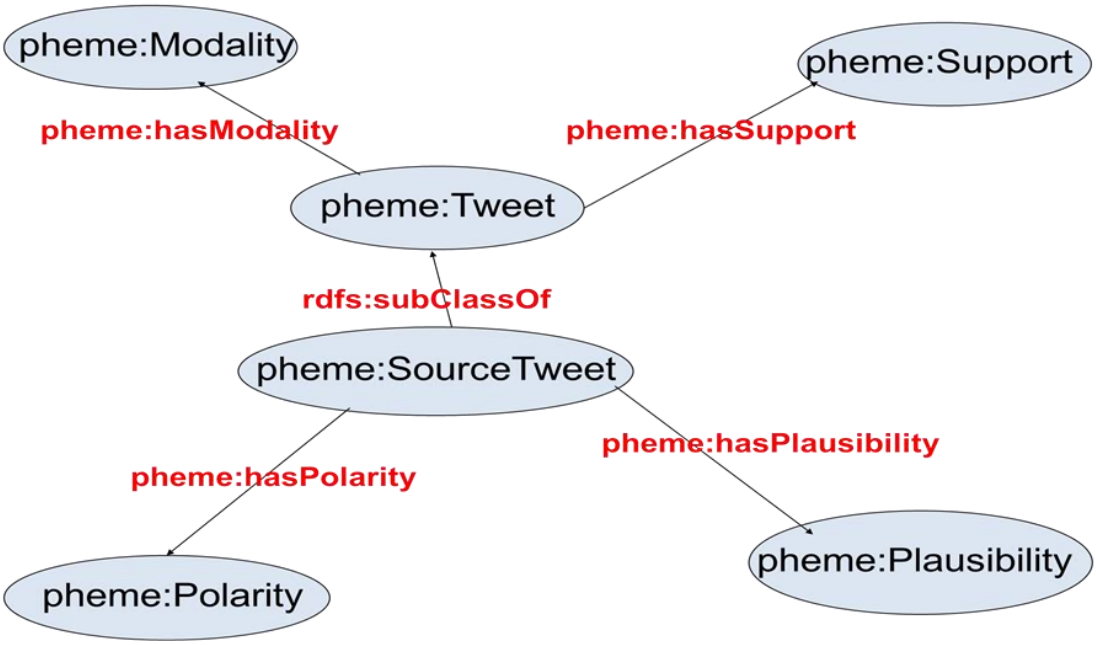

Figure 8: The properties associated with the SourceTweet class: *hasModality, hasSupport, hasPolarity, hasPlausability*.

And finally in Figure 9, below, we display the properties of the ResponseTweet class: *hasModality*, *hasSupport* (Support), *hasRespondType* (ResponseType). In this figure only specific for the Pheme ontology classes and properties are given.

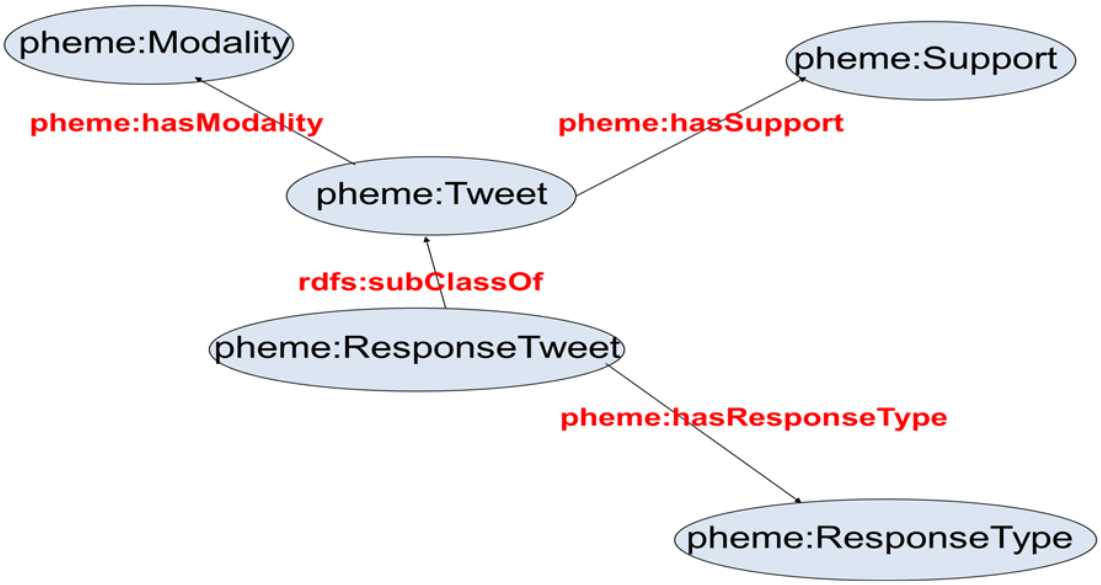

Figure 9: the properties of the ResponseTweet class

2.1.3 Conclusion and current Work

In the figures above we have given a sketch of the actual organisation of the Pheme ontology, which in this first round of development is mainly responding to the Pheme deliverable D2.1 "Qualitative Analysis of Rumours, Sources, and Diffusers across Media and Languages".

We have shown how the Pheme model has been integrated in the PROTON upper model, and how other ontologies (for example the LivePost ontology) can be imported/integrated. As the project is going on, we will add further specifications resulting from the work on the use cases in WP7 (*Veracity Intelligence in Patient Care*) and WP8 (*Digital Journalism Use Case*). For each case, we have already developed an ontology model. For WP7 a model based on the ATC (Anatomical Therapeutic Chemical) classification system[8], and for WP8, a model based on the IPTC (International Press Telecommunications Council) standard. These models will be integrated in the Pheme ontology.

Current work is also on populating the Pheme ontology with annotated data from the use cases partners.

[8] This system is „used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties." (http://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System?oldformat=true)

3 Multilingual Pre-processing

Pheme aims to work over large volumes of data, both streaming and in batches, taken from informal sources such as social media and web forums. Automatic rumour identification and categorisation requires sophisticated models which interpret and process data, including text, in a standardised manner, in order to be able to first make and then apply generalisations. To this end, the text data that Pheme works with needs to be interpreted and formalised as accurately as possible, to enable the best possible rumour detection.

3.1 Introduction

While schemata and tools are widely available for processing English text, these are typically learned from and evaluated over newswire or other formally-structured text genres. Pheme operates over less-constrained, more varied data, which is not always so well handled by these existing tools. In addition, general resources for Bulgarian and German are more limited than those for English, and social media adapted tools for these two languages are almost non-existent or just emerging.

In this section of D2.2, we identify some sources of representative web data, and then report on the state-of-the art in language pre-processing tools for each language with regard to web/social media text. The key research output is the identification of tools to be reused in Pheme, based on analysis of existing multilingual pre-processing tools.

3.2 English

For English, we have framed evaluation and tool selection around a web text toolkit consisting of the following parts: language ID, part-of-speech tagging, named entity recognition, entity extraction and parsing. Each stage is either specifically adapted to web text, or the best state-of-the-art third party tool empirically selected.

3.2.1 Data Sources

The level of structure found in text on the web varies greatly. At one extreme are articles from formal news source, such as the Wall Street Journal (WSJ)[9], which has a very confined bias on multiple levels. Its style is dictated by editorial guidelines and press body rules (e.g. Associated Press style); in addition, articles about business mergers and stock prices follow a set structure, with a sequence of paragraphs covering set topics and in a specific order. In terms of latent bias, the Wall Street Journal articles used for a large part of Natural Language Processing (NLP) research were written in the 1980s and 1990s, where the authors are typically working-age, middle-class, white American men (Eisenstein, 2013). This has an impact on the word

[9] A description of this corpus is available at https://catalog.ldc.upenn.edu/LDC93S6A

choice, grammar, and conceptual structures used. Articles are usually at least a few hundred words long. Finally, each article is proof-read and then reviewed by an editor before publication.

At the other extreme are microblog texts, i.e. those found on social media sites like Tumblr and Twitter. These texts have less latent bias than the WSJ; the only requirement for posting on sites like this is access to the internet and some basic skill in using it. Accordingly, the background of microblog authors is broader; for example, in the US, of all the microblogging platforms, Twitter attracts the largest proportion of African-American users, comprising around 40% of accounts (Søgaard, 2014). Style on Twitter is varied, though often retains strong structural elements which are consistent per-user (Hu, 2013). The result is a perfusion of styles which don't really occur in existing language corpora used in NLP.

3.2.2 Evaluation over tweets

The huge variation in Twitter web text, and its stark difference from conventional newswire (e.g. WSJ) text, makes it one of the hardest types of web text to process (Derczynski, 2015). As this is where the most text processing problems emerge, we use performance over tweets for our evaluation.

For our evaluation, we use existing Twitter datasets distributed as part of prior work, for ease of comparison. For the construction of new tools, we use: (1) a combination of these annotated Twitter datasets; (2) a four-year archive of tweets held by USFD that were gathered as part of the Trendminer project[10]; and (3) an ongoing collection of tweets gathered by USFD in real-time.

3.2.3 Language ID

Since microblog content is strongly multilingual, language identification needs to be performed prior to other linguistic processing, in order to ensure that the correct later components are run. Named entity recognition depends on correct language identification. Not only do names have different spellings in different languages, but critically, the surrounding syntactic and lexical context is heavily language-dependent, being filled with e.g. names of jobs, or prepositions.

TextCat (van Noord, 1997) and the Cybozu language detection library (Shuyo, 2010) rely on n-gram frequency models to discriminate between languages, relying on token sequences that are strong language differentiators. Information gain-based langid.py (Lui, 2012) uses n-grams to learn a multinomial event model, with a feature selection process designed to cope with variations of expression between text domains. Both TextCat and langid.py have been adapted for microblog text, using human-annotated data from Twitter. The former adaptation works on a limited set of languages; the latter on 97 languages (Preotiuc 2012).

We evaluated system runs on the ILPS TextCat microblog evaluation dataset (Carter, 2013). Results are given in Table 16, with the Twitter-specific versions marked

[10] See http://www.trendminer-project.eu/ for more details

"twitter". Microblog-adapted TextCat performed consistently better than the other systems.

The adapted version of TextCat has a slightly easier task than that for langid.py because it expects only five language choices, whereas the adapted langid.py is choosing labels from a set of 97 languages. The latter assigned a language outside the five available to 6.3% of tweets in the evaluation set. The adapted langid.py performed worse than the generic version. The results are quite close for some languages, and so if an approximate 6% improvement could be made in these cases, the Twitter-adapted version would be better. Although language identification is harder on tweets than on longer texts, its performance is sufficiently high to inform reliable choices in later stages.

Our novel contribution here is a broad empirical evaluation of language ID tools, including those used in the previous TrendMiner pipeline. This evaluation tells us how to get the best results out of the Pheme web text toolkit.

3.2.4 Part-of-speech tagging

It is well known that part-of-speech tagging for tweets is much harder than for news text. Of particular importance is accuracy at whole-sentence level, which is required for effective event extraction and argument/event association.

Traditional off-the-shelf part of speech tagging tools perform poorly with twitter text. As a result, a number of taggers have emerged specifically for twitter text. A leading one of these comes from USFD (Derczynski 2013). This tagger has been updated with extra training data, annotated as part of Pheme, to improve performance. In addition, other datasets (Owoputi 2013, Hovy 2014) have been converted into the rich Penn TreeBank tagset using techniques developed at USFD (notably "vote-constrained bootstrapping", VCB), allowing their use with the tagger. This rich tagset is sufficiently detailed to capture some morphological and dependency information, which is not the case for the tagsets used in competing twitter POS taggers; these use the simplified CMU (Gimpel 2011) or Google tagsets (Petrov 2011).

Table 1: Sentence tagging, token tagging; Pheme performances compared with other approaches

Tagger	Token accuracy	Sentence accuracy
Stanford	73%	2%
Ritter	85%	9%
Derczynski (2013)	89%	20%
Pheme	92%	26%

3.2.5 Named entity recognition

Named entity recognition (NER) is a critical Information Extraction (IE) task, as it identifies which snippets in a text are mentions of entities in the real world. It is a pre-requisite for many other IE tasks, including coreference resolution and relation extraction. NER is difficult on user-generated content in general, and in the microblog genre specifically, because of the reduced amount of contextual information in short

messages and a lack of curation of content by third parties (e.g. that done by editors for newswire).

A plethora of named entity recognition techniques and systems is available for general full text. For Twitter, some approaches have been proposed but they are mostly still in development, and often not freely available. In the remainder of this section, we evaluate and compare a mixture of Twitter-specific and general purpose NER tools. We want to eliminate the possibility that poor NER on tweets is systematic – that is, related to some particular approach, tool, or technology.

For our analyses of generic NER systems, we chose those that take different approaches and are immediately available as open source. The first system we evaluate is ANNIE from GATE version 8, which uses gazetteer-based lookups and finite state machines to identify and type named entities in newswire text. The second system is the Stanford NER system (Finkel 2005), which uses a machine learning-based method to detect named entities, and is distributed with CRF models for English newswire text.

Of the NER systems available for Twitter, we chose that of Ritter (2011), who take a pipeline approach performing first tokenisation and POS tagging before using topic models to find named entities, reaching 83.6% F1 measure. In addition to these, we also include a number of commercial and research annotation services, available via Web APIs and a hybrid approach, named NERD-ML, tailored for entity recognition of Twitter streams, which unifies the benefits of a crowd entity recognizer through Web entity extractors combined with the linguistic strengths of a machine learning classifier.

The commercial and research tools which we evaluate via their Web APIs are Lupedia, DBpedia Spotlight, TextRazor, and Zemanta. DBpedia Spotlight and Zemanta allow users to customize the annotation task, hence we applied the following settings: 1) DBpedia Spotlight={confidence=0, support=0, spotter=CoOccurrenceBasedSelector, version=0.6}; and 2) Zemanta={markup limit:10}.

Their evaluation was performed using the NERD framework (Rizzo 2012) and the annotation results were harmonized using the NERD ontology. The NERD core ontology provides an easy alignment with the classes used for this task.

Table 2: Evaluation of NER systems applied to tweets

	Per entity type F1				Overall F1		
System	**Location**	**Misc**	**Organisation**	**Person**	**P**	**R**	**F1**
ANNIE	40.23	0.00	16.44	24.81	36.14	16.29	22.46
DBpedia Spotlight	46.06	6.99	19.44	48.55	34.70	28.35	31.20
Lupedia	41.07	13.91	18.92	25.00	38.85	18.62	25.17
NERD-ML	**61.94**	23.73	**32.73**	**71.28**	52.31	**50.69**	**51.49**
Stanford	60.49	**25.24**	28.57	63.22	**59.00**	32.00	41.00
Stanford-Twitter	60.87	25.00	26.95	64.00	54.39	44.83	49.15
Textrazor	36.99	12.50	19.33	70.07	36.33	38.84	37.54
Zemanta	44.04	12.05	10.00	35.77	34.94	20.07	25.49

We can see that conventional tools (i.e., those trained on newswire) perform poorly in this genre, and thus microblog domain adaptation is crucial for good NER. However, when compared to results typically achieved on longer news and blog texts, state-of-

the-art in microblog NER is still lacking good results. Consequently, there is a significant proportion of missed entity mentions and false positives.

There are a number of reasons for the low results of the systems on the Ritter dataset. Partly, this is due to the varying annotation schemes and corpora on which the systems were trained. The annotation mapping is not perfect: for example, we mapped Facility to Organisation, but some systems will be designed to represent Facility as Location, in some or all cases. Similarly, some systems will consider MusicArtist as a kind of Person, but in our mapping they are Misc, because there are also bands. All this means that, as is common, such a comparison is somewhat imperfect and thus the comparative results are lower than those usually reported in the system-specific evaluations. It should also be noted that this is also a single-annotator corpus, which has implications for bias that make it hard to discern statistically significant differences in results

The kinds of entities encountered in microblog corpora are somewhat different from those in typical text. We subjectively examined the entities annotated as people, locations and organisations in the microblog corpora and the CoNLL NER training data[11]. For people, while those mentioned in news are often politicians, business leaders and journalists, tweets talk about sportsmen, actors, TV characters, and names of personal friends. The only common type is celebrities. For locations, news mentions countries, rivers, cities – large objects – and places with administrative function (parliaments, embassies). Tweets on the other hand discuss restaurants, bars, local landmarks, and sometimes cities; there are rare mentions of countries, often relating to a national sports team, or in tweets from news outlets. Finally, for organisations, the news similarly talks about organisations that major in terms of value, power or people (public and private companies and government organisations) while tweets discuss bands, internet companies, and sports clubs. Tweets also have a higher incidence of product mentions than the news genre, occurring in around 5% of messages.

That the entities occurring in tweets are different from those in newswire makes it hard for systems to tag them correctly. Ratinov and Roth (2009) point out that given the huge potential variety in named entity expressions, unless one has excellent generalised models of the context in which entities are mentioned, it becomes very difficult to spot previously unseen entities. This applies to gazetteer-based approaches in particular, but also to statistical approaches. Twitter is well-known as being a noisy genre, making it hard even for systems with perfect models of entity context to recognise unseen NEs correctly. For example, in newswire, person entities are often mentioned by full name, preceded by a title, constitute the head of a prepositional phrase, or start a sentence, and are always correctly capitalised. They are often followed by a possessive marker or an extended description of the entity. This kind of linguistic context is well-formed and consistent, possibly having stability bolstered by journalistic writing guidelines. In contrast, person entities in tweets are apparently semi-stochastically capitalised, short, and occur in a variety of contexts (Bontcheva 2014) – including simply as exclamation. This is a hostile tagging environment, where one will suffer if one expects the cues learned from heavily structured newswire to be present.

[11] CoNLL stands for Conference on Computational Natural Language Learning. This conference series has been organizing regularly so-called shared tasks. One such shared task was on Named Entity Recognition (NER), see http://www.cnts.ua.ac.be/conll2003/ner/.

In Pheme so far, we have identified suitable systems and adaptations for NER over tweets, and found that it is a much tougher task than over typical web text, as expected. We also identified a specific novel adaptation based on existing techniques. Using the Stanford system's feature extraction, we used an adapted version of Conditional Random Field **(**CRF), namely with passive-aggressive weight updates, which are less subject to noise, something rife in web text. This led to instant performance increases (Derczynski, 2014). After this, we also applied gazetteer-based filtering using an intelligent post-editing component that picked up common ambiguous or hard-to-recognise terms and specifically concentrated on those in a second pass after initial entity recognition.

This discriminative post-editing reclassified borderline entity matches with F1 of 83.8, including finding 92.7% of missed person-type entities in a benchmark set. We examine two types of classification error: false positives (spurious) and false negatives (missed). False positives occur most often where non-person entities are mentioned. This occurred with mentions of organisations (Huff Post), locations (Galveston) and products (Exodus Porter). Descriptive titles were also sometimes mis-included in person names (Millionaire Rob Ford). Names of persons used in other forms also presented as false positives (e.g. Marie Claire – a magazine). Polysemous names (i.e. words that could have other functions, such as a verb) were also mis-resolved (Mark). Finally, proper nouns referring to groups were sometimes mis-included (Haitians). Despite these errors, precision almost always remained higher than recall over tweets. We use in-domain training data, and so it is unlikely that this is due to the wrong kinds of person being covered in the training data – as can sometimes be the case when applying tools trained on newswire.

False negatives often occurred around incorrect capitalisation and spelling, with unusual names, with ambiguous tokens and in low-context settings. Both omitted and added capitalisation gave false negatives (charlie gibson, or KANYE WEST). Spelling errors also led to missed names (Russel Crowe). Ambiguous names caused false negatives and false positives; our approach missed mark used as a name, and the surname of Jack Straw. Unusual names with words typically used for other purposes were also not always correctly recognised (e.g. the Duck Lady, or the last two tokens of Spicy Pickle Jr.). Finally, names with few or no context words were often missed (Video: Adele 21., and 17-9-2010 Tal al-Mallohi, a 19-).

In summary, with Pheme we have investigated a broad range of entity recognition systems, found the best-performing ones, and then discovered specific adaptations that improve practical performance on web/twitter text.

3.2.6 Entity linking

Microblog named entity linking (NEL) is a relatively new, underexplored task. Recent Twitter-focused experiments uncovered problems in using state-of-the-art entity linking in tweets. This is again largely due to lack of sufficient context to aid disambiguation. Others have analysed Twitter hashtags and annotated them with DBpedia entries to assist semantic search over microblog content. Approaches based on knowledge graphs have been proposed, in order to overcome the dearth of context problem, with some success. Given the shortness of microtext, correct semantic interpretation is often reliant on subtle contextual clues, and needs to be combined

with human knowledge. For example, a tweet mentioning iPad makes Apple a relevant entity, because there is the implicit connection between the two.

To evaluate entity linking for English, we constructed a small corpus of 182 tweets and expert-sourced (i.e. a crowdsourced model, like CrowdFlower -- CF[12], though using only approved expert annotators) entity disambiguation results for the entities mentioned within. The corpus was taken from the overlap of the Ritter POS and NER datasets, to enable maximum reusability in future research. Evaluation was performed using direct strict matches only: that is, any entity extent mistakes or disambiguation errors led to an incorrect mark. Future evaluation measures may be interested in both lenient NER evaluation (with e.g. overlaps being sufficient) as well as lenient entity disambiguation grading, through e.g. allowing entities connected through skos:exactMatch or even owl:sameAs to be credited as correct. However, for this work, we stay with the strictest measure.

Results are given below. Note that the recall here is not reliable, as we have "downsampled" the set of URIs returned so that all fall within DBpedia, and DBpedia does not contain references to every entity mentioned within the dataset. In addition, due to the specifics of the CrowdFlower NEL interface, annotators could not indicate missed entities for which no candidates have been assigned, even if such exist in DBpedia. Therefore, no result is placed in bold in this dataset. We perform this downsampling so that we may find the simplest possible base for comparison.

Table 3: Results of the evaluation of NEL for English

Name	Precision	Recall	F1
DBpedia Spotlight	7.51	27.18	11.77
YODIE (simple)	36.08	**42.79**	39.15
YODIE (news-trained)	**67.59**	33.95	**45.20**
Zemanta	34.16	28.92	31.32

Amongst the systems compared here, YODIE[13] performed best, which should be attributed at least partly to it using Twitter-adapted pre-processing components.

One major source of mistakes is capitalisation. In general, microblog text tends to be poorly formed and typical entity indicators in English such as midsentence capitalisation are often absent or spuriously added to non-entities. To demonstrate this point, we compare the proportional distribution of lower, upper and sentence-case words in a sentence with the tokens which are subject to false positives or false negatives, in the table below. It can be seen that the majority of false positives are in sentence case, and the largest part of false negatives are in lowercase. Conversely, almost no false positives are in lower case. In both cases, very different proportions are mislabelled from the underlying distributions of case, indicating a non-random effect. Tweet text is often short, too, and sometimes even human annotators did not have enough contexts to disambiguate the entities reliably. For example, one candidate tweet mentioned "at Chapel Hill". While "Chapel Hill" is almost certainly a location, it is a very common location name, and not easy to disambiguate even for a human. Indeed, annotators disagreed enough about this tweet for it to be excluded from the final data.

[12] See http://www.crowdflower.com/ for more details.
[13] YODIE is he Ontology-based IE system developed in part at USFD: https://gate.ac.uk/applications/yodie.html

Also, many named entities are used in unconventional ways. For example, in the tweet "Branching out from Lincoln park after dark ... Hello "Russian Navy", it's like the same thing but with glitter! ", there are no location or organisation entities. Note that Russian Navy is collocated with glitter, which is a little unusual (this could be determined through language modelling). It refers in fact to a nail varnish colour in this context, and is undisambiguable using DBpedia or many other general-purpose linked entity repositories. Lincoln Park After Dark is another colour, and compound proper noun, despite its capitalisation – it does not refer to the location, Lincoln Park.

The general indication is that contextual clues are critical to entity disambiguation in microblogs, and that sufficient text content is not present in the terse genre. As usual, there is a trade-off between precision and recall. This can be affected by varying the number of candidate URIs, where a large number makes it harder for a system to choose the correct one (i.e. reduces precision), and fewer candidate URIs risks excluding the correct one entirely (i.e. reduces recall).

3.2.7 Dependency and Shallow parsing

Determining sentence structure in tweets is critical to effective event and argument extraction. It also can rely on having good part-of-speech sequence extraction, as mentioned above.

We have evaluated approaches to dependency parsing and to shallow parsing. Dependency parsing aims to build a tree-like structure of words in the sentence, describing which concepts act or depend on others. This tree becomes a representation of the relationships between each word in the sentence, conveying its overall meaning. Shallow parsing works at a simpler level, identifying contiguous sequences of words that have a specific unified function – for example, verb phrases or noun phrases.

For dependency parsing, we used the evaluation of Malt provided in Foster (2011). This compared WSJ-trained resources with those based on a custom twitter corpus as input, and adapted Malt to re-train using bootstrapped data. This achieved an 11% performance increase.

Table 4: Evaluation of the Malt parser applied to twitter text

Tagger	Accuracy
Malt (basic)	67.64%
Malt (adapted)	78.67%

Regarding shallow parsing, we compared the performance of OpenNLP (Baldridge, 2005) and Ritter's system (2011). Results are given below. In this case, Ritter's system built upon Twitter-specific adaptations to part-of-speech-tagging and added Brown cluster information to help work around the orthographic variance prevalent in twitter text. This achieved a 22% error reduction in parsing accuracy over twitter text compared to the non-adapted system. Further improvement in POS tagger accuracy will lead to even higher performance.

Table 5: Evaluation of 2 chnukers applied to twitter text

System	Accuracy

OpenNLP	0.839
Ritter	0.875

3.3 Bulgarian

Bulgarian is a less-resourced language in the area of processing social media. Thus, some adaptation of the basic Bulgarian NLP pipeline available at ONTO was needed. And also a Bulgarian Tweet Corpus was created.

3.3.1 Types of Annotation for the BulgarianTweet Corpus

The annotation for the BulgarianTweet Corpus was done in two ways: using a basic NLP pipeline and by its extended part with URIs from DBpedia. First, 16 308 tweets with no special unifying topic were processed with the basic NLP pipeline. The aim was the identification of sources of errors and the resulting adaptation of the pipeline for tweets processing.

Then, tweets related to *Bank Crisis in July 2014* were semi-automatically annotated with DBpedia URIs with the help of the CLaRK system[14]. This sub-corpus consisted of 1150 tweets, containing 24721 tokens. And we had 1410 named entities, annotated with DBpedia URIs.

The Bulgarian part of DBpedia was used for the annotation. The various spellings of the names were extracted from the URIs (where possible). A regular grammar was constructed in CLaRK system. A frequency list of the tokens in the tweets was constructed and used as a filter for minimizing the size of the grammar (it was also used as a source of additional spellings of the existing NEs). For the missing NEs records Wikipedia was used directly. New URIs were created and classified for the missing NEs.

The annotated data contained ambiguities and missing links. Thus, it was post-edited manually.

3.3.2 The NLP pipeline

The processing pipeline for Bulgarian is provided by IICT-BAS[15]. Its components are implemented in Java. The Bulgarian processing module includes a combination of: rule-based components (tokenization, sentence splitting, lemmatization), hybrid components (POS tagging) and statistical components.

Since Bulgarian is an analytical language with rich word inflection, the task of POS tagging becomes more complex. It is better defined as morphosyntactic annotation due to the high number of grammatical features encoded in the tagset.

[14] See http://www.bultreebank.org/clark/

[15] See http://www.lrec-conf.org/proceedings/lrec2012/summaries/829.html

The BulTreeBank NLP Pipeline has the following evaluation metrics on the BulTreeBank data:

POS tagging: 97.98 %
Lemmatization: 95.23 %
Dependency Parsing: 92.9 %

The adaptation steps included the following processes: Orthographic normalization, Tokenization, out-of-vocabulary elements (OOV), parts-of-speech tagging (POS tagging), Named entity recognition (NER) and Dependency parsing.

The tokenization showed 2 % of erroneous result on the Bulgarian Tweet Corpus. It addressed the following problems: Latin and Cyrillic letters, Punctuation, Emoticons, Links. The most problematic issue remained the erroneously written Latin letters in Cyrillic tokens.

The OOV showed 1 % of erroneous result. The module had to face the following problems, which are typical for the colloquial and concise nature of the tweets:

- Specific contractions of words (<tok>кардио</tok>, 'cardio' instead of 'кардиологичен', cardiologic)
- Errors of various kinds – spelling, grammatical, etc. (<tok>фзатвора</tok> (the preposition is written together with the noun and it is also written as pronounced, not as codified) instead of 'в затвора', in prison)
- Colloquial sublanguage which can show also dialectal nuances (<tok>тея</tok> (highly colloquial) instead of 'тези', these; <tok>Наааайс</tok> 'transliteration from the English word 'nice'' instead of 'хубаво', nice).

The POS tagging showed 2,26 % mistagged tokens. The mistagging was either due to errors in the tokenization module, or due to unknown words.

The adaptation went in the following directions: tokenizer was adapted to process also emoticon, punctuation and unusual letter and number combinations. However, not all of them get a POS tag. Thus, the dependency parsing operates very well on typical texts and fails on texts with no POS tags.

Figure 10 shows an example of a tweet, annotated with dependency relations (the visualization is from the CLaRK system). It says: "Well, what is going on with First Investment Bank?". Here the contraction к'во (from какво, 'what') was analyzed correctly.

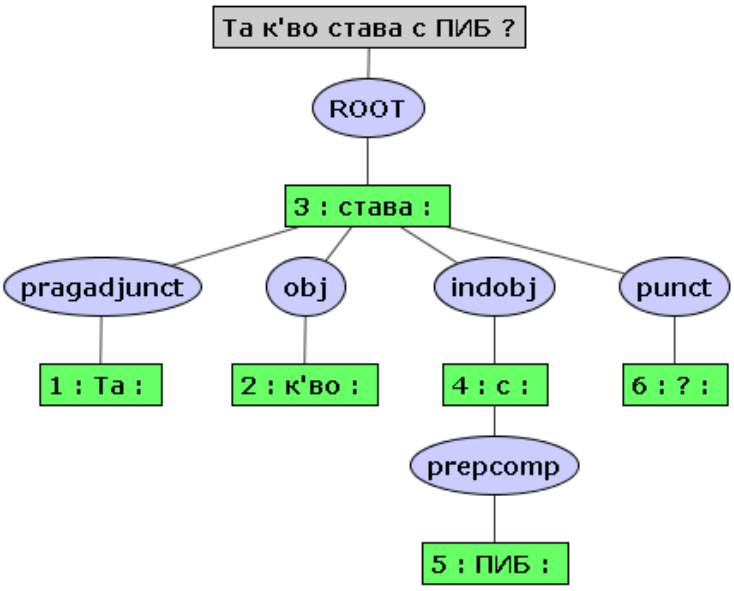

Figure 10: Example of a Bulgarian tweet annotated with dependency relations.

Real evaluation of the pipeline on tweet data was not performed, since our Tweet corpus was annotated only with DBpedia URIs as gold standard. At the moment hashtags are not analysed.

3.4 German

The partner USSAR (University of Saarland) conducted a study on the performance of off-the-shelve POS taggers when applied to Internet texts in German, and investigated easy-to-implement methods to improve tagger performance. The main findings were that extending a standard training set with small amounts of manually annotated data for Internet texts can lead to a substantial improvement of tagger performance, which can be further improved by using a method to automatically acquire training data. For the evaluation of the approach, a manually annotated corpus of Internet forum and chat texts, consisting of 24 000 tokens, has been created. In the further course of the project, we will use directly German Pheme text data. Results of this work are described in (Horbach et al., 2014), which we summarize here.

3.4.1 Data sources

Two complementary types of Internet text – forum posts from the Internet cooking community www.chefkoch.de and the Dortmund Chat Corpus (Beißwenger, 2013) – were selected in order to cover a range of phenomena characteristic of Internet-based communication. As mentioned above, we are currently now building a German corpus of typically Pheme text data, which was not available at the beginning of the project.

3.4.2 Tagset

As to be expected (see the related comments in sections 3.2.1 and 3.3.1 for the English and Bulgarian Twitter corpus data) this corpus data contain some language phenomena that are not properly covered by tagsets that have been developed mainly for processing newspaper corpora, so for example the standard STTS tagset[16] for German, such as emoticons, so called "action words" in inflective form, URLs and various kinds of contractions. In order to account for the most frequent of those phenomena, an extended version of STTS has been proposed by Bartz et al. (2014).

Two tags have been added to this extension of STTS in order to capture errors made by the writers. The tag ER-RAW is assigned when a token should be part of the following token, i.e. if the writer inserted an erroneous whitespace; the tag ERRTOK is a tag for the opposite case when the writer joined two words that should be in fact separated.

3.4.3 Annotation

11 658 tokens from the Dortmund Chat Corpus and 12 335 tokens from randomly chosen posts from the Chefkoch corpus have been manually annotated with POS information. Prior to annotation, the data has been automatically tokenized.

Systematic errors during this process have been cleaned up manually. Annotators were asked to ignore token-level errors like typos or grammatical errors whenever possible. For instance, when the conjunction "dass" was erroneously written "das", they should annotate KOUS even though "das" as a correct form can only occur as ART, PRELS or PDS.

3.4.4 Corpus Analysis

The two subcorpora vary considerably not only in general linguistic properties like average sentence length (10:5 tokens for forum, 5:9 for chat) but even more so in the frequency with which POS tags, especially the non-standard tags occur. While the forum data only contain 3% of nonstandard tags, chat contains 11.2% of those new tags, thus clearly calling for adapted processing tools. And in the chat corpus only 60.0% of all sentences are covered by the traditional STTS tagset.

3.4.5 The Experiment on Part-of-Speech Tagging

2 distinct methods are tested, using for this also 3 different statistical POS taggers. The 2 methods consisted in 1) Extend a standard newspaper-based training corpus

[16] STTS stands for Stuttgart-Tübingen-Tagset. See http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html

with data drawn from automatically tagged Internet texts applying a technique proposed by Kübler and Baucom (2011); and 2) extend the training corpus with small portions of manually annotated forum and chat texts.

Results show that while the first approach leads to minor improvements of tagger performance, it is outperformed by a large margin by the second approach – even if only very few additional training sentences are added to the training corpus. A small further improvement can be obtained by combining the two approaches.

The result has been verified through the application of three different statistical parsers: TreeTagger (Schmid, 1994), Stanford Tagger (Toutanova et al., 2003) and TnT (Brants, 2000).

A larger gain (13:4 % for the best performing TnT tagger) in performance of the taggers applied to the chat corpus data can be obtained by adding small amounts of manually annotated.

This is the result that is guiding our work on applying POS tagging to the German corpus data of Pheme to be delivered soon to the technological partners.

4 Conclusions

In this deliverable, we have first presented the actual state of modelling the types of information objects that are globally needed in Pheme. In the second part on the linguistic processing, we have evaluated a range of approaches and systems for processing user-generated content against purpose-built datasets constructed as part of the project, and identified both which approaches and systems perform best on low-context, high-variance web text (i.e. Tweets and Chats), as well as conducted a thorough error analysis to pave the way for future work in the area.

List of Abbreviations

LOD – Linked Open Data
NE – Named Entities
NER – Named Entity Recognition
NEL– Named Entity Linking
NLP – Natural Language Processing
PROTON – PROTo Ontology
OOV – Out-of-vocabulary Elements
OWL – Web Ontology Language
OWL 2DL – Web Ontology Language Description Language
OWL 2RL – Web Ontology Language Rule Language
POS tagging – Part-of-speech tagging

Bibiliography and references

Sofia Angeletou, Matthew Rowe, Harith Alani, 2011. Modelling and Analysis of User Behaviour in Online Communities. In Proceedings of the International Semantic Web Conference (ISWC).

Baldridge et al., 2005. The OpenNLP project. http://opennlp.apache.org/

Thomas Bartz, Michael Beißwenger, and Angelika Storrer, 2014. Optimierung des Stuttgart-Tubingen-Tagset für die linguistische Annotation von Korpora zur internet-basierten Kommunikation: Phanomene, Herausforderungen, Erweiterungsvorschläge. Zeitschrift für germanistische Linguistik, 28(1):157–198.

Michael Beißwenger, 2013. Das Dortmunder Chat-Korpus. Zeitschrift für germanistische Linguistik 41(1):161–164.

Bontcheva and Derczynski, 2014. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text: slides. Figshare, doi http://dx.doi.org/10.6084/m9.figshare.1003767.

Thorsten Brants, 2000. TnT – a statistical part-of-speech tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing, Association for Computational Linguistics.

Carter et al., 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Journal of Language Resources and Evaluation, 47(1).

Derczynski et al., 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In Proceedings of RANLP.

Derczynski and Bontcheva, 2014. Passive-Aggressive Sequence Labeling with Discriminative Post-Editing for Recognising Person Entities in Tweets. In Proceedings of EACL.

Derczynski et al., 2015. Analysis of Named Entity Recognition and Linking for Tweets. Information Processing and Management, 51(2).

Eisenstein, 2013. What to do about bad language on the internet. In Proceedings of NAACL.

Finkel et al., 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of ACL.

Foster et al., 2011. #hardtoparse: POS Tagging and Parsing the Twitterverse. In Proceedings of AAAI.

Gimpel et al., 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In Proceedings of ACL.

Andrea Horbach, Diana Steffen, Stefan Thater, Manfred Pinkal, 2014. Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In Proceedings of the 12th Konvens Conference.

Hovy et al., 2014. When POS datasets don't add up: Combatting sample bias. In Proceedings of LREC.

Hu et al, 2013. Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. In Proceedings of ICWSM.

Sandra Kübler and Eric Baucom, 2011. Fast domain adaptation for part of speech tagging for dialogues. In Proceedings of RANLP 2011.

Lui and Baldwin, 2012. langid.py: An off-the-shelf language identification tool. In Proceedings of ACL.

Owoputi et al., 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In Proceedings of NAACL.

Petrov et al., 2011. "A universal part-of-speech tagset." arXiv preprint arXiv:1104.2086.

Preotiuc et al., 2012. Trendminer: An architecture for real time analysis of social media text. In Proceedings of the workshop on real-time analysis and mining of social streams.

Ratinov and Roth, 2009. Design Challenges and Misconceptions in Named Entity Recognition. In Proceedings of CoNLL.

Ritter et al., 2011. Named entity recognition in tweets: an experimental study. In Proceedings of EMNLP.

Rizzo and Troncy, 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In Proceedings of EACL.

Scerri, K. Cortis, I. Rivera, and S. Handschuh, 2012. Knowledge Discovery in Distributed Social Web Sharing Activities. In Proceedings of the #MSM2012 Workshop.

Helmut Schmid, 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing

Shuyo, 2010. Language detection library for java. http://code.google.com/p/language-detection/

Søgaard, 2014. Semantic parsing for the 99%. In Proceedings of the Workshop on semantic annotation and processing.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics

Van Noord, 1997. TextCat–An Implementation of a Text Categorization Algorithm.