The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods

Behrang QasemiZadeh† **and Anne-Kathrin Schumann**‡
†DFG Collaborative Research Centre 991, Heinrich-Heine University Düsseldorf
‡Department of Applied Linguistics, Translation and Interpreting, Saarland University
zadeh@phil.hhu.de, annek_schumann@gmx.de

Abstract

This paper introduces the *ACL Reference Dataset for Terminology Extraction and Classification*, version 2.0 (ACL RD-TEC 2.0). The ACL RD-TEC 2.0 has been developed with the aim of providing a benchmark for the evaluation of term and entity recognition tasks based on specialised text from the computational linguistics domain. This release of the corpus consists of 300 abstracts from articles in the ACL Anthology Reference Corpus, published between 1978–2006. In these abstracts, *terms* (i.e., single or multi-word lexical units with a specialised meaning) are manually annotated. In addition to their boundaries in running text, annotated terms are classified into one of the seven categories *method*, *tool*, *language resource (LR)*, *LR product*, *model*, *measures and measurements*, and *other*. To assess the quality of the annotations and to determine the difficulty of this annotation task, more than 171 of the abstracts are annotated twice, independently, by each of the two annotators. In total, 6,818 terms are identified and annotated, resulting in a specialised vocabulary made of 3,318 lexical forms, mapped to 3,471 concepts. We explain the development of the annotation guidelines and discuss some of the challenges we encountered in this annotation task.

Keywords: Terminology, Term Extraction, Term Classification, Entity Recognition, Evaluation Corpus, Language Resource.

1. Introduction

Terminology mining methods are the cornerstone of modern information systems. These methods are concerned with the automatic analysis of *languages for special purposes*—for example, to facilitate knowledge acquisition from text, to enhance the interoperability when communicating knowledge between communities or across languages, to provide information summaries, and so on. One way or another, these methods deal with the extraction of lexical items known as *terms*.

Linguistically, a term is a *lexical unit* that carries a *specialised meaning* in a particular *context* (e.g., see Faber and Rodríguez (2012)). According to their lexical forms, terms are either *simple* (made of one word/token) or *complex* (i.e., multi-word units). The extraction of terms has been addressed under various names, different conditions (e.g., by restrictions applied to their input and output), and for colourful purposes. The most familiar examples, perhaps, are methods for automatic term extraction (ATR), key-phrase extraction, and entity extraction—see QasemiZadeh (2015, chap. 3) for an explanation of similarities and differences between these methods. The evaluation of all these methods, however, has been tackled in a similar fashion.

The black-box, data-driven evaluation process—as initiated in the series of message understanding conferences (MUC) and still used in semantic evaluation workshops (SemEval), etc.—is the dominant methodology for assessing and comparing the performance of term extraction methods.[1] MUC-like evaluation systems consist of two major components. The first component is a *gold* dataset: that is, a collection of manually annotated text. The second component is a collection of performance measures (e.g., precision, recall, . . .). Using the performance measures, an extraction method is then assessed by comparing its output to the annotations in the 'gold standard'. This paper introduces the ACL RD-TEC 2.0 that has been developed to serve as the first component of this kind of evaluation methodology.

Previously introduced by QasemiZadeh and Handschuh (2014), the *ACL RD-TEC* provides manual annotations for a set of more than 80,000 lexical units extracted from the ACL Anthology Reference Corpus—also known as the ACL ARC (Bird et al., 2008). These lexical units are manually annotated either as valid, or invalid terms. Valid terms are then further categorised as *technology* and *non-technology* terms.[2] Hence, the ACL RD-TEC is suitable for the evaluation of ATR and term classification methods.

In this second release, the ACL RD-TEC dataset is extended in two ways. First, instead of providing annotations for a list of isolated lexical units, 300 abstracts from the ACL ARC are fully annotated for the terminology they contain. Two annotators[3] (with expertise in computational linguistics) carefully read these abstracts, and in accordance with a detailed set of guidelines, they marked lexical boundaries for all the terms they encountered. Secondly, we extend the term categorisation scheme. Instead of technology and non-technology terms, terms are semantically grouped into 7 co-hyponym categories (see Table 1). An example of an annotated abstract is shown in Listing 1. Consequently, apart from ATR and term classification, the second release of the ACL RD-TEC is suitable for the evaluation of entity recognition methods.

In the remainder of this paper, Section 2 further explains the motivation behind the development of this resource. In Section 3, we describe the process of developing the annotation guidelines. In Section 4, we report changes that are observed in the inter-annotators' agreement during this

[1] See Lehnert et al. (1994) for a detailed description.

[2] ACL RD-TEC is available through ELDA (Behrang QasemiZadeh, 2014).

[3] That is, the authors of this paper.

#	Category	Description	Example
1	Technology and Method	Terms referring to practical tasks, processes, and solutions in NLP	machine translation, speech recognition, . . .
2	Tool and Library	Names of implemented (actualised) methods and libraries	OpenNLP, Sphinx, . . .
3	Language Resource	Components of NLP solutions containing linguistic knowledge	lexicon, parallel corpus, . . .
4	Language Resource Product	actualised (instances of) language resources	WordNet, Brown Corpus, . . .
5	Models	terms refer to *encoded* linguistic knowledge	language model, translation model,. . .
6	Measures and Measurements	mainly components of evaluation systems used for measuring and measurement processes	BLEU, Precision,. . .
0	Other	Any category other than listed above (e.g., theories, formalism, linguistic entities, . . .); this category is likely to embrace very specific terms.	target language, orthographical variation, . . .

Table 1: Semantic categories in the ACL RD-TEC 2.0.

Listing 1: Example of an annotated abstract.

```
<?xml version="1.0" encoding="UTF-8" standalone="
    ↪ no"?>
<Paper acl-id="P05-2016">
<Title>Dependency-Based Statistical Machine
    ↪ Translation</Title>
<Section>
<SectionTitle>Abstract</SectionTitle>
<S>We present a <term class="tech">Czech-English
    ↪ statistical machine translation system</
    ↪ term> which performs <term class="tech">
    ↪ tree-to-tree translation</term> of <term
    ↪ class="other">dependency structures</term>
    ↪ .</S>
<S>The only <term class="lr">bilingual resource</
    ↪ term> required is a <term class="lr">
    ↪ sentence-aligned parallel corpus</term>.</
    ↪ S>
<S>All other <term class="lr">resources</term>
    ↪ are <term class="other">monolingual</term>
    ↪ .</S>
<S>We also refer to an <term class="measure(ment)
    ↪ ">evaluation method</term> and plan to
    ↪ compare our <term class="other">system's
    ↪ output</term> with a <term class="measure(
    ↪ ment)">benchmark system</term>.</S>
</Section>
</Paper>
```

procedure. In Section 5, we report the current state of the corpus and its manual annotations. We conclude in Section 6 by discussing challenges met during the process of annotating this corpus.

2. Why AC RD-TEC?

Searching publications and browsing language resource directories, one finds that several datasets comparable to the ACL RD-TEC already exist—for example the GENIA (Kim et al., 2003) or the CRAFT corpora (Bada et al., 2012). Therefore, it is likely that a curious mind points to these resources and asks 'why do we need ACL RD-TEC?'.
Firstly, for computational linguists, the use of resources in domains other than computational linguistics—for example, the GENIA corpus in the domain of molecular biology—is hindered by an obstacle: the minimal prerequisite knowledge that is required to understand this literature (not to mention the specialised discourse and style of writing in these domains). This understanding of text, perhaps, is essential to enable a computational linguist to first comprehend and then describe a linguistic phenomenon. Hence, text mining in a specialised domain is often conducted by a team that includes experts in the domain under investigation, or computational linguists who have specialized training (as best exemplified by research teams conducting biotext mining). Conducting text analyses and lexicography in domains other than computational linguistics, therefore, may not be the first, best choice for computational linguists.
Secondly, existing corpora and resources for benchmarking terminology extraction methods largely ignore the temporal aspect of the development of terms (and thus knowledge and technologies). In this sense, the ACL ARC and consequently the ACL RD-TEC 2.0 are unique resources: The annotated corpus contains scientific texts published during more than three decades.[4] Therefore, this resource can be employed as a valuable asset in research investigating the history of science (e.g., as in Schumann and QasemiZadeh (2015b)), apart from trending tasks such as *trend analysis* (e.g., Mariani et al. (2014)).
Thirdly, in the development of the ACL RD-TEC, an important emphasis has been put on the transparent development of the resource. We believe that the difficulty of the problem of identifying terms (i.e., their lexical boundaries in running text as well as their semantic connotations) has been largely overlooked. We maintain that annotating terms and building specialised vocabularies is a much harder task than building resources for similar tasks such as for named entity recognition. Deciding whether a lexical unit (i.e., either a word or a phrase) in a given context is a term or not is not a straightforward decision. Decisions of this type are likely to be influenced by presupossitions and extra-linguistic factors such as text length, etc.
To investigate this problem, apart from keeping an inventory of the changes applied during the development of the guidelines, more than 171 abstracts in the ACL RD-TEC are annotated by 2 experts. By publishing these annotations separately in one resource,[5] we aim to provide a re-

[4] Note that the next release of the ACL ARC will contain literature from 1965 to 2015—that is, 50 years of accumulated knowledge in the domain of computational linguistics.

[5] And, by encouraging the annotation of these abstracts by ad-

source that for addressing the requirements for qualitative assessments of the difficulties encountered when developing specialised vocabularies.

3. Development of the Annotation Guidelines

The guidelines used for the development of the ACL RD-TEC 2.0 can be found in Schumann and QasemiZadeh (2015a). Initially, these guidelines were drafted based on those previously used for the development of the first release of the dataset (see QasemiZadeh (2014)). Apart from providing general information about the task (such as a concise theoretical background, information about the text being annotated, examples, . . .), the guidelines consist of a number of concrete rules and criteria which are categorised as *semantic*, *linguistic*, and *formal*.

The *semantic criteria* elaborate an organisation of terms into several categories of concepts (see Table 1). *Formal criteria* spell out rules for deciding about the boundaries of terms, one of the major challenging tasks in the annotation process—for example, rules for dealing with term abbreviation sequences (such as "machine translation (MT)"), rules for deciding whether adjectival modifiers should be included into the term span as well as rules describing how to split longer noun phrases containing several term candidates (e.g. "TREC 2003 and TREC 2004 QA tracks" and "automatic evaluation of machine translation and document summarization"). Finally, *linguistic criteria* enumerate linguistic characteristics of terms (of which the most important one is that annotations are restricted to nominals).

In order to develop and refine these three criteria, we employed an iterative process of annotating a small, fixed set of abstracts (Figure 1). In the first iteration, a set of 10 abstracts was selected randomly. Independently of each other, we annotated these abstracts. In turn, using a tool that was developed to compare annotations against each other, conflicting annotations (regarding the boundaries of terms and their semantic categorisations) were found and discussed by the annotators. As the outcome of this discussion, the guidelines were elaborated and new rules were added, when necessary. Using the new, refined guidelines, the same procedure was repeated: a set of 22 abstracts was annotated from scratch, and the resulting annotations were compared to refine the guidelines. This iterative process was repeated until we agreed that the guidelines were sufficient and could not be improved (i.e., *four* iterations altogether). Interested readers can find additional information about the evolution of the guidelines by browsing the document history appended to the guidelines document.

For asserting annotations in the abstract, we simply use text editors and tags in a 'semi-XML' format. The sanity of manual annotations is then checked using a self-coded Java tool and in turn the annotated abstract files are converted to a valid XML format as shown in Listing 1.

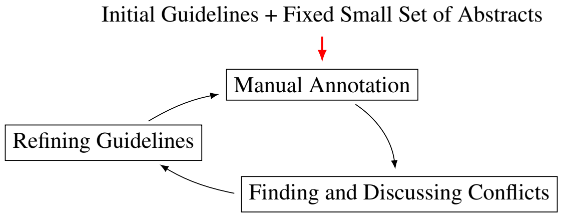
Initial Guidelines + Fixed Small Set of Abstracts
Manual Annotation
Finding and Discussing Conflicts
Refining Guidelines

Figure 1: The iterative process of developing and refining the annotation guidelines: Given a small, fixed set of abstracts, the initial guidelines were refined through an iterative process to achieve higher inter-annotators' agreement.

			Total#	
#i	**Boundary**	**Overall**	**A$_1$**	**A$_2$**
1	0.528	0.471	121	95
2	0.709	0.652	129	105
3	0.711	0.595	512	349
4	0.755	0.635	413	337

Table 2: Changes in the annotators' agreement achieved through the iterative refinement of the guidelines. Agreement is assessed using the F-scores when all the annotations from different abstracts are consolidated in one list. #i denotes the number of iterations; Total# shows the number of annotations asserted by the first (i.e., A$_1$) and the second annotator (A$_2$) in each iterations. Note that for #i=3 and 4, 12 additional abstracts are added.

4. Inter-Annotators' Agreement During the Development of the Guidelines

In this section, we report the statistics regarding the inter-annotators' agreement (IAA) during the process of developing the guidelines. Since we cannot give a clear estimation of the number of possible terms—and their length—in the text being annotated (given the fact that many terms are multi-word units), chance-corrected IAA measures such as Cohen's κ cannot be applied in this context. Simply put, in order to use chance-corrected measures such as Cohen's κ, the set of entities in the annotation task must be known prior to the task.[6]

To assess the annotators' agreement, we calculate the *F-score* measure as follows:

$$F\text{-}score = \frac{2 \times r \times p}{r + p}, \qquad (1)$$

in which p (i.e., precision) and r (i.e., recall) are computed as $p = \frac{|tp|}{|tp|+|fp|}$, and $r = \frac{|tp|}{|tp|+|fn|}$. In these equations, $|tp|$ denotes the number of lexical items that are annotated as *terms* by both of the annotators; $|fp|$ denotes the number of lexical items that are annotated as terms only by the first annotator; similarly, $|fn|$ is the number of lexical items that are annotated as terms only by the second annotator.[7]

To count the number of the true and false positives as well as the false negatives, in the first instance, we use exact matches for comparing the term boundaries and their semantic classes. As an example, consider the text snippet:

ditional members of the community, for example, the authors of the abstracts themselves.

[6] Assuming that the number of possible terms in text is very large, the hypothetical probability of chance agreement, in fact, approaches 0 (i.e., $p_e \to 0$). Note that from an ultimate perspective, any combination of words in a sentence can be a term. If $|W|$ is the number of words in a given sentence S, potentially there are $2^{|W|}$ terms in S. Thus, the claim that $p_e \to 0$ seems reasonable.

[7] Evidently, in this context *true negatives* are not defined.

#i	Boundary		Overall	
	μ	σ	μ	σ
1	0.49	0.039	0.423	0.04
2	0.692	0.044	0.638	0.032
3	0.724	0.04	0.602	0.039
4	0.741	0.023	0.63	0.018

Table 3: Changes in the annotators' agreement achieved through the iterative refinement of the guidelines using the *averaged F-scores*. F-scores are computed per abstract and then are averaged as a figure of merit of the agreement between annotators. #i denotes the number of iterations; μ and σ denote the arithmetic mean and the standard deviation, respectively.

. . . for robust natural language processing in . . .

If the first annotator marks "robust natural language processing" and the second annotator marks "natural language processing" as a term, then we count these annotations as unmatched (i.e., either as *false positive* or *false negative* depending on which annotator is taken as a reference[8]). Since the boundaries of the two terms do not match, their semantic classes will also remain unmatched.

Given the description above, we compute and report the F-scores (see Equation 1) for the assessment of the annotators' agreement for:

(a) *Only the term boundaries*: In the example above, using this measure, it is verified whether both annotators mark "statistical natural language processing" as a term.

(b) *Overall*: Both term boundaries and their assigned semantic classes are matched. Using this measure, we check whether the annotation are indeed identical on both categories. Thus, the reported F-score for item (a) sets the upper bound limit for the *Overall* performance.

These measures are reported for each of the four iterations. Table 2 lists the computed F-scores when annotations from all the abstracts are consolidated in one list (to cancel bias that may result from the difference in the length and number of sentences in abstracts). Similar results are reported in Table 3 where the computed F-scores are averaged over the set of abstracts in the pilot annotation task. That is, the F-Scores are first computed per abstract file and then they are averaged to report the overall performance. These averaged numbers may better represent the performances of annotators when they deal with different topics in the corpus (e.g., machine translation, anaphora resolution, etc.). Table 4 details the number of annotated terms and their distribution in the envisaged semantic classes.

To further detail the performance of annotators in the subtask of assigning semantic classes to terms, we report the inter-annotators' agreements on this sub-task by limiting the input only to the subset of terms that both annotators have identified—that is, an intersection of resulting annotations for term boundaries (see Table 4, numbers embraced in parentheses). For this case, however, we use Cohen's κ as an indication of the inter-annotators' agreement. For the obtained set of annotations in each iteration, κ is measured using:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad (2)$$

where the observed agreement probability p_o is given by:

$$p_o = \frac{1}{|T|} \sum_{c \in C} a_c, \qquad (3)$$

where a_c is the number of terms assigned to the same semantic class $c \in C$ by both annotators, and T is the total number of annotated terms ($|T|$). Likewise, the probability of agreement by chance (p_e) is given by:

$$p_e = \sum_{c \in C} p_c \times p'_c, \qquad (4)$$

where p_c and p'_c are the probabilities of assigning semantic category c to a term by the first and the second annotator, respectively.[9] Table 5 shows the computed κ for each iteration; similar to Table 2 and 3, these numbers are computed for (a) when all terms (of identical span) are collected from different abstracts and compared in one go, and (b) when κ scores are computed per abstract and then averaged.

As shown, while IAA are "mostly" satisfactory (F-Score above 0.70 for term boundaries and F-Score and κ around 0.60 in semantic class assignment), they are still away from complete agreement. Despite our efforts during this procedure—that is, extensive discussions as well as the elaboration of the criteria in the guidelines both for term boundaries and semantic classes—we were not able to always resolve conflicts and choose one of the suggested values for term boundaries or semantic classes as the "correct answer". As mentioned earlier, presuppositions and the level of annotators' expertise in topics they deal with, as well as factors such as text length, the style of writing (and even errors in the original text) may well play a role in the motivation of disagreement. These factors certainly must be also taken into consideration when developing and evaluating term recognition and classification methods.

To emphasise and investigate the difficulty of the task, we report the *self-agreement* for annotating a fixed set of abstracts using the same guidelines. In a first session, each annotator performed the annotations for 10 abstracts; in a second session a day after this first session, the annotator repeated the annotation task for these 10 abstracts. Table 6 reports the agreement between annotations obtained from these two sessions for each of the annotators. As Table 6 shows, even a highly trained annotator cannot reach a 100% agreement on annotating term boundaries and semantic classes with himself (herself)—that is, annotations produced by one annotator for the same text in two different sessions are not identical. The reported self-agreement in Table 6, perhaps, can be viewed as a realistic upper bound baseline for the assessment of IAA scores.

[8] Note that choosing the reference annotator does not affect the computed F-scores.

[9] Note that *model* and *measure(ment)* have been added as semantic categories only after the 3rd iteration.

	Iterations							
	1		2		3		4	
Category	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
LR	7 (5)	7 (6)	6 (5)	11 (10)	41 (31)	49 (45)	28 (19)	27 (25)
LR-Prod	0 (0)	0 (0)	0 (0)	0 (0)	6 (5)	8 (7)	1 (1)	0 (0)
Measure(ment)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	51 (31)	35 (23)
Model	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	15 (12)	34 (31)
Tech	45 (21)	48 (26)	47 (28)	38 (28)	141 (83)	85 (72)	137 (100)	99 (91)
Tool	1 (0)	1 (0)	1 (1)	1 (1)	6 (2)	16 (16)	4 (4)	2 (2)
Other	68 (31)	39 (25)	75 (49)	55 (44)	318 (185)	191 (166)	177 (116)	140 (111)
Total#	121 (57)	95 (57)	129 (83)	105 (83)	512 (306)	349 (306)	413 (283)	337 (283)

Table 4: Distribution of the assigned semantic classes per annotator per iteration. Numbers in parentheses show the statistics only for the subset of text spans that are marked as term by both annotators.

		Averaged κ per Abst.	
#iteration	κ	μ	σ
1	0.76	0.584	0.156
2	0.787	0.774	0.048
3	0.608	0.56	0.107
4	0.67	0.656	0.057

Table 5: IAA using Cohen's κ only for the subset of terms with identical boundaries, when all annotated terms are compared disregarding the abstract file they appear in, and (b) when the κ score is measured per abstract and averaged to report the performance.

Annotator	**Boundary**		**Overall**		**κ Category**	
	μ	σ	μ	σ	μ	σ
1	0.881	0.093	0.854	0.006	0.889	0.019
2	0.823	0.029	0.713	0.036	0.562	0.138

Table 6: Self-agreement in the annotation task: The same set of abstracts are annotated using the same guidelines in two sessions that were apart a day. Reported numbers are computed similar to those reported in Table 3 and Table 5 (i.e., they are averaged over the abstracts).

5. Statistics for the Annotated Abstracts

The ACL RD-TEC 2.0 consists of 300 annotated abstracts. To choose these abstracts, we employ stratified random sampling. Abstracts that could be extracted automatically using the ParsCit tool (Councill et al.,) (about 8500 abstracts) from the raw text files are grouped by their year of publication (i.e., 1965–2006). For each year from 1978, a number of abstracts are selected randomly while maintaining each year's proportional share in the overall number of publications. The resulting 300 abstracts are then segmented into sentences[10] in which OCR[11] and segmentation errors are corrected manually. The result is a corpus of 1,384 sentences and 33,216 tokens.

Category	**A_1**	**A_2**	**IAA_o**
LR	125 (95)	127 (109)	0.69
LR-Prod	22 (18)	14 (13)	0.667
Measure(ment)	119 (90)	119 (90)	0.63
Model	91 (77)	232 (202)	0.44
Tech	720 (589)	713 (629)	0.738
Tool	50 (43)	42 (41)	0.826
Other	1459 (1108)	1169 (936)	0.678
Total#	2586 (2020)	2416 (2020)	0.727

Table 7: Statistics for the 171 abstract files that are annotated twice. A_1 and A_2 denote annotators. Parenthesised numbers show the count of commonly identified terms boundaries (i.e., both annotators agree on term boundaries, but not necessarily on semantic classes). The IAA_o shows the F-Score resulting from the identification of terms with identical boundaries and semantic classes.

Among these 300 abstracts, 171 are annotated twice (i.e., independently by each of the annotators). Hence, the published dataset contains 471 annotated files: 179 abstracts annotated only once (by one of the annotators), 171 abstracts annotated by the first annotator, and the same set of 171 abstracts annotated by the second annotator. Note that these abstracts are labelled in a way that they can be grouped by the year of their publication and annotator; therefore, one may fuse or remove annotations from the files according to the evaluation context.

The double-annotated 171 abstracts consist of 817 sentences and 19,476 tokens. In total, in these 171 abstracts, both annotators have identified 2,020 common term boundaries which results in an F-score of 0.808 for the identification of term boundaries. The overall F-score performance (i.e., annotated terms with the same boundary and the same semantic class as explained in Section 4) is 0.727.

Table 7 details statistics about the number of terms marked by each annotator and their agreement for the assignment of semantic classes. For instance, as shown in the first row of Table 7, the first annotator identifies 125 terms of semantic class *LR* where the second annotator identifies 127 LR terms. Amongst the 125 terms that the first annotator marked as LR, only 95 are also identified as valid term by the second annotator.These 95 terms, however, are not necessarily classified as LR by the second annotator as indicated in the table (see IAA_o). As shown, the highest agreement has been achieved in annotating terms that refer to specific software packages(i.e., category *tool*) while the least reliable annotations are for the category *model*.

We conclude this section by reporting detailed statistics for various linguistic entities and annotations in this release of the corpus (see Table 8).[12] The corpus is freely available to

[10] Using OpenNLP pre-trained sentence splitting (`https://opennlp.apache.org/`.)

[11] Optical Character Recognition.

[12] The corpus can be browsed on-line, see `http://pars.ie/`

				Semantic Categories						
	Abstract	**Sentence**	**Term**	LR	LR-Prod	Measure(ment)	Model	Tech	Tool	Other
Annotator 1	189	900	2857	145	27	131	102	790	54	1608
Annotator 2	282	1301	3961	231	33	183	328	1124	83	1979
Total#	471	2201	6818	376	60	314	430	1914	137	3587
Unique#	300	1384	4849	276	47	218	338	1314	94	2562

Table 8: Statistics for the ACL RD-TEC 2.0: Shown are the number of annotated units, and the distribution of the semantic classes assigned to terms by each annotator (first and second row). The row marked by 'total#' shows the sum of all entities, including identical annotations that are manually asserted by both annotators. The 'unique#' row, however, shows the numbers for uniquely asserted entries in the dataset. As mentioned earlier, the 4,849 term annotations yield to a specialised vocabulary of size 3,318.

download from the LINDAT/CLARIN Infrastructure.

6. Discussion

Building on knowledge gathered from previous experience in the development of the first release of the ACL RD-TEC, the ACL RD-TEC 2.0 annotates terms in 300 abstracts and classifies them into 7 categories of concepts. Half of the abstracts in this release of the corpus are double-annotated and they are published independently with the (additional) goal of providing materials for analysing agreement between annotators and investigating difficulties in constructing terminological resources.

The ACL RD-TEC 2.0, by all means, is still far away from an ideal resource for benchmarking terminology extraction and classification. However, as a wise man once said, 'perfection is the enemy of good'; despite all its shortcomings, the ACL RD-TEC 2.0 brings us a step closer towards building a resource of familiar materials for computational linguists who are interested in topics such as terminology extraction, history of science, and automatic content analysis of scholarly publications.

The presented work and the annotation scheme can be extended in many ways: Semantic categories can be refined and extended, many layers of annotations must be added to the corpus (e.g., term variants must be marked, morphosyntactic information must be inserted, and so on). Apart from these classic considerations, during the development of this resource, we have identified a number of linguistic phenomena that have been less discussed in terminology publications, particularly in the context of language resource development.

We have found contextual variation and ellipsis to constitute a source of challenge and conflict during the annotation task. In many cases, authors shorten terms and use just the head of the nominal group instead of the whole term. In particular, this is often the case in anaphoric contexts (e.g., as in "...our method ...", "...proposed algorithm..."). Deciding whether to annotate these structures or not was a source of controversy during our annotation task. While we have omitted this type of structure from this release of the corpus, more elaboration on this topic is certainly an interesting avenue for future work. We encountered a similar problem when annotating coordinate structures (e.g., as in "...information extraction and retrieval...").

In addition, study of systematic polysemy in specialised text can be an interesting future research. In the context of our annotation task, for example, we noticed that terms that are often classified as "method" are also used to denote the specific "problem" (or "task') that they are designed for. In many occasions, annotation of these terms were a cause of disagreement between annotators. Lastly, we recognise that the annotation of linguistic units other than nominals (e.g. verbs and adjectives) is also important for building a comprehensive resource.

7. Acknowledgements

The work described in this paper was partially funded by the DFG (Deutsche Forschungsgemeinschaf) through the Cluster of Excellence "Multimodal Computing and Interaction" and the "Collaborative Research Centre 991 (CRC 991): *The Structure of Representations in Language, Cognition, and Science*". We are grateful to Prof. Elke Teich and Prof. Laura Kallmeyer for supporting the development and dissemination of this resource.

8. Bibliographical References

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W., Cohen, K., Verspoor, K., Blake, J., and Hunter, L. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, May. ELRA.

Councill, I. G., Giles, C. L., and Kan, M.). Parscit: an open-source CRF reference string parsing package. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, May. ELRA.

Faber, P. and Rodríguez, C. I. L. (2012). Terminology and specialized language. In Pamela Faber, editor, *A Cognitive Linguistics View of Terminology and Specialized Language*, volume 20 of *Applications of Cognitive Linguistics*, pages 9–33. Walter de Gruyter.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(Suppl 1):i180–i182.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1994). Evaluating an information ex-

`lr/acl_rd-tec`.

traction system. *Journal of Integrated Computer-Aided Engineering*, 1(6).

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014). Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of LREC'14*, Reykjavik, Iceland, May. ELRA.

QasemiZadeh, B. and Handschuh, S. (2014). The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the Computerm'14 Workshop*, Dublin, Ireland, August. ACL.

QasemiZadeh, B., (2014). *The ACL RD-TEC: Annotation Guideline (Ver 1.0).* National University of Ireland.

QasemiZadeh, B. (2015). *Investigating the Use of Distributional Semantic Models for Co-Hyponym Identification in Special Corpora*. Ph.D. thesis, National University of Ireland, Galway, July.

Schumann, A.-K. and QasemiZadeh, B., (2015a). *The ACL RD-TEC Annotation Guidelines*. Saarland University and National University of Ireland, ver. 2.6 edition. Available from `http://pars.ie/publications/papers/pre-prints/acl-rd-tec-guidelines-ver2.pdf`.

Schumann, A.-K. and QasemiZadeh, B. (2015b). Tracing Research Paradigm Change Using Terminological Methods: A Pilot Study on 'machine translation' in the ACL Anthology Corpus. In *Proceedings of TIA'15*, Granada, Spain, November. CEUR Workshop Proceedings.

9. Language Resource References

Behrang QasemiZadeh. (2014). *ACL RD-TEC: A Reference Dataset for Terminology Extraction and Classification Research in Computational Linguistics*. Behrang QasemiZadeh, distributed via ELRA, Terminology, 1.0, ISLRN 699-305-362-089-6.