Towards comprehensive syntactic and semantic annotations of the clinical narrative

Daniel Albright,[1] Arrick Lanfranchi,[1] Anwen Fredriksen,[1] William F Styler IV,[1] Colin Warner,[2] Jena D Hwang,[1] Jinho D Choi,[3] Dmitriy Dligach,[4] Rodney D Nielsen,[1,5] James Martin,[3] Wayne Ward,[3] Martha Palmer,[1] Guergana K Savova[4]

[1]Department of Linguistics, University of Colorado, Boulder, Colorado, USA
[2]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Department of Computer Science University of Colorado, Boulder, Colorado, USA
[4]Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[5]Department of Computer Science and Engineering, University of North Texas, Texas, USA

Correspondence to
Dr Guergana K Savova, Boston Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02114, USA; Guergana.Savova@childrens.harvard.edu

Received 3 September 2012
Revised 27 December 2012
Accepted 28 December 2012
Published Online First
25 January 2013

To cite: Albright D, Lanfranchi A, Fredriksen A, *et al. J Am Med Inform Assoc* 2013;**20**:922–930.

ABSTRACT

Objective To create annotated clinical narratives with layers of syntactic and semantic labels to facilitate advances in clinical natural language processing (NLP). To develop NLP algorithms and open source components.

Methods Manual annotation of a clinical narrative corpus of 127 606 tokens following the Treebank schema for syntactic information, PropBank schema for predicate-argument structures, and the Unified Medical Language System (UMLS) schema for semantic information. NLP components were developed.

Results The final corpus consists of 13 091 sentences containing 1772 distinct predicate lemmas. Of the 766 newly created PropBank frames, 74 are verbs. There are 28 539 named entity (NE) annotations spread over 15 UMLS semantic groups, one UMLS semantic type, and the Person semantic category. The most frequent annotations belong to the UMLS semantic groups of Procedures (15.71%), Disorders (14.74%), Concepts and Ideas (15.10%), Anatomy (12.80%), Chemicals and Drugs (7.49%), and the UMLS semantic type of Sign or Symptom (12.46%). Inter-annotator agreement results: Treebank (0.926), PropBank (0.891–0.931), NE (0.697–0.750). The part-of-speech tagger, constituency parser, dependency parser, and semantic role labeler are built from the corpus and released open source. A significant limitation uncovered by this project is the need for the NLP community to develop a widely agreed-upon schema for the annotation of clinical concepts and their relations.

Conclusions This project takes a foundational step towards bringing the field of clinical NLP up to par with NLP in the general domain. The corpus creation and NLP components provide a resource for research and application development that would have been previously impossible.

INTRODUCTION

With hospitals and governments around the world working to promote and implement widespread use of electronic medical records, the corpus of generated-at-the-point-of-care free-text to be processed for relevant phenotypic information continues to grow. Although some types of notes (eg, radiology or pathology reports) are often formulaic in nature, others (eg, clinical notes (CN)) afford doctors the freedom to create medical documentation with all the expediency, nuance, and implications present in natural language. For example, it is still challenging for language processing technologies to reliably discover the experiencer of a clinical event (patient, family member, or other), the level of certainty associated with an event (confirmed, possible, negated) as well as textual mentions that point to the same event. We describe our efforts to combine annotation types developed for general domain syntactic and semantic parsing with medical-domain-specific annotations to create annotated documents accessible to a variety of methods of analysis including algorithm and component development. We evaluate the quality of our annotations by training supervised systems to perform the same annotations automatically. Our effort focuses on developing principled and generalizable enabling computational technologies and addresses the urgent need for annotated clinical narratives necessary to improve the accuracy of tools for extracting comprehensive clinical information.[1] These tools can in turn be used in clinical decision support systems, clinical research combining phenotype and genotype data, quality control, comparative effectiveness, and medication reconciliation to name a few biomedical applications.

In the past decade, the general natural language processing (NLP) community has made enormous strides in solving difficult tasks, such as identifying the predicate-argument structure of a sentence and associated semantic roles, temporal relations, and coreference which enable the abstraction of the meaning from its surface textual form. These developments have been spurred by the targeted enrichment of general annotated resources (such as the Penn Treebank (PTB)[2]) with increasingly complex layers of annotations, each building upon the previous one, the most recent layer being the discourse level.[3] The emergence of other annotation standards (such as PropBank[4] for the annotation of the sentence predicate-argument structure) has brought new progress in the annotation of semantic information. The annotated corpora enable the training of supervised machine learning systems which can perform the same types of annotations automatically. However, the performance of these systems is still closely tied to their training data. The more divergent the test data are, such as the gap between newswire and clinical narrative data, the lower the performance.

In an effort to capture some of this progress and allow rapid alignment of clinical narrative data with other corpora, tools, workflows, and community adopted standards and conventions, we incorporated several layers of annotation from the general domain into our clinical corpus, each optimized for information extraction in a specific area. For syntactic information, all notes were annotated following the PTB model,[5–11] and to capture predicate-argument structure, a PropBank[4] annotation layer was created. To capture the complex semantic types present in the clinical narrative, we used the Unified Medical Language System (UMLS) Semantic Network schema of entities.[12 13] The established nature of these annotations provides a valuable advantage when porting existing algorithms or creating new ones for information extraction from the clinical narrative and moving towards semantic processing of the clinical free text.

To our knowledge, this is the first clinical narrative corpus to include all of these syntactic and semantic layered annotations, making these data a unique bridge for adapting existing NLP technology into the clinical domain. We built several NLP components based on the annotations—a part-of-speech (POS) tagger, constituency parser, dependency parser, and semantic role labeler (SRL)—and did indeed find the expected performance improvements. These components have been contributed to the open-source, Apache clinical Text Analysis and Knowledge Extraction System (cTAKES).[14] To facilitate research on the corpus, the corpus described herein will be made available through a data use agreement with the Mayo Clinic.[i]

BACKGROUND

We embarked on the task of creating a corpus of layered annotations and developing NLP components from it as part of a bigger project focused on building a question answering system for the clinical domain, the Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ).[15 16] Because the information sought through questions posed by end users could potentially span the entire domain of medicine, one of the requirements for the system is comprehensive information extraction, which in turn entails comprehensive semantic processing of the clinical narrative.

Within the clinical domain, there are only a handful of annotated clinical narrative corpora. Ogren and colleagues[17] developed a corpus of 160 CN annotated with the UMLS semantic group of Disorders. Each Disorder entity mention was mapped to a UMLS concept unique identifier (CUI). The Clinical E-Science Framework (CLEF) corpus[18] is annotated with information about clinical named entities (NEs) and their relations as well as with temporal information about the clinical entities and time expressions that occurred in the clinical narrative. It consists of CN, radiology reports, and histopathology reports together with associated structured data. The entity annotations are normalized to the UMLS semantic network. The relations are of types *has_target*, *has_finding*, *has_indication*, *has_location*, and *modifies*. Temporal expressions follow the TimeML standard[19]; temporal relations are of types *before*, *after*, *overlap*, and *includes*. Unfortunately, the corpus has not been released to the research community.

For the 2010 i2b2/VA NLP challenge, a corpus of CN was annotated for concepts, assertions, and relations.[20] The medical concepts were of types *problem*, *test*, and *treatment*. The assertions for each *problem* concept described whether the concept was 'present,' 'absent,' 'possible,' 'conditional,' 'hypothetical,' or 'associated with someone else.' The annotated relations were between pairs of concepts within a sentence, one being a *problem*. Types of relations were *treatment is given for the problem*, *treatment is not given because of the problem*, *treatment worsened the problem*, *test revealed the problem*, and *problem indicates another problem*. The corpus consists of 826 documents and is available to researchers through data use agreements.

The Bioscope Corpus[21] consists of annotations of medical and biological texts for negation, speculation, and their linguistic scope with the goal of facilitating the development and evaluation of systems for negations/hedge detection and scope resolution.

The MiPACQ clinical corpus presented here differs from previous work in the layers of annotations, their comprehensiveness and adherence to community adopted conventions and standards—syntactic annotations following PTB guidelines, predicate-argument semantic annotations following PropBank guidelines, and UMLS entity semantic annotations. Thus, the layered structure allows the development of interoperable enabling technologies critical for semantic processing of the clinical narrative. We developed and evaluated such technologies to demonstrate the utility of the corpus and the expected performance gains.

METHODS

Corpus

The MiPACQ clinical corpus consists of 127 606 tokens of clinical narrative, taken from randomly selected Mayo Clinic CN, and Mayo Clinic pathology notes related to colon cancer. All notes have been completely anonymized. In comparison, the Wall Street Journal (WSJ) PTB we use here for training contains 37 015 sentences, with 901 673 word-tokens. Research was conducted under an approved Institutional Board Review protocol from the Mayo Clinic.

Annotation layers

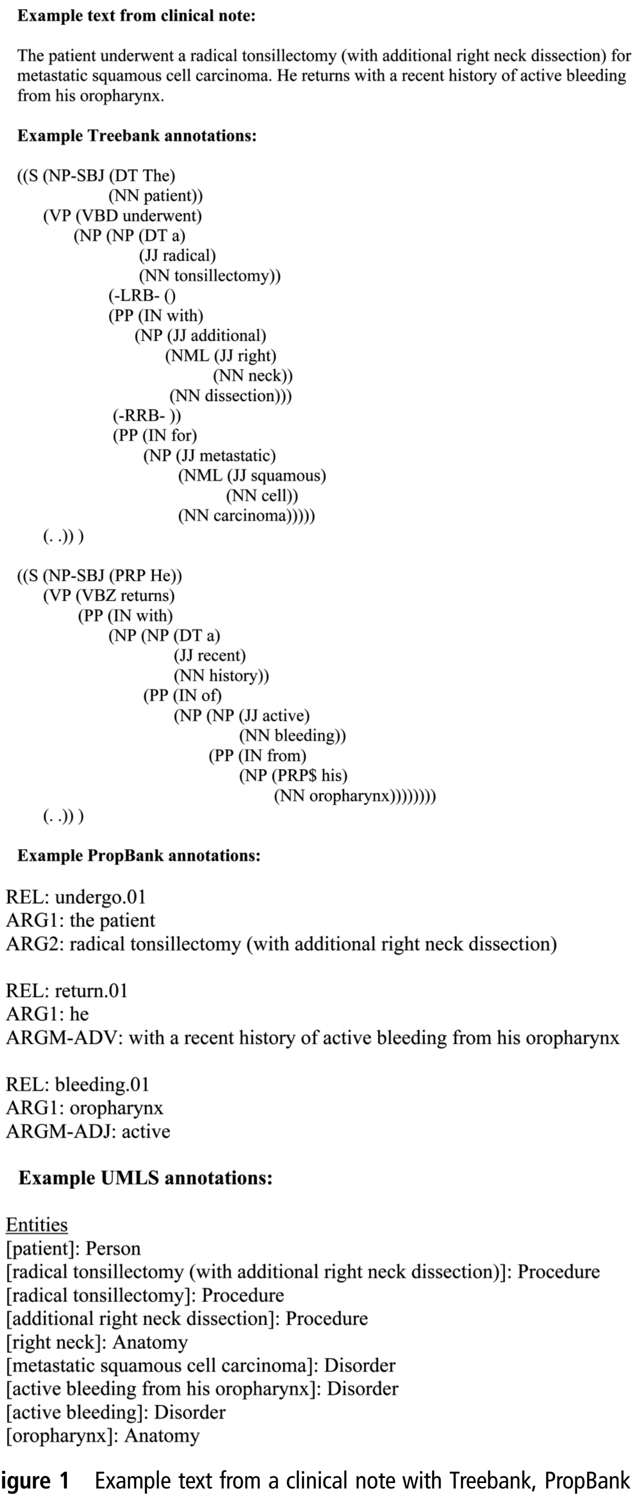
For each layer of annotations, we developed explicit guidelines.[22–24] Figure 1 is used as a running example.

Treebank annotations

Treebank annotations consist of POS, phrasal and function tags, and empty categories (see below), which are organized in a tree-like structure (see figure 1). We adapted Penn's POS Tagging Guidelines, Bracketing Guidelines, and all associated addenda, as well as the biomedical guideline[8] supplements. We adjusted existing policies and implemented new guidelines to account for differences encountered in the clinical domain. PTB's supplements include policies for spoken language and biomedical annotation; however, CN contain a number of previously unseen patterns that required the refinements in the PTB annotation policies. For example, fragmentary sentences like 'Coughing up purulent material.' which are common in clinical data, are now annotated as S with new function tag, –RED, to mark them as reduced and given full subject and argument structures. Under existing PTB policies these are annotated as top-level FRAG and lack full argument structure. Figure 2 presents examples of Treebank changes. Additionally, current PTB tokenization was fine-tuned to handle certain abbreviations more accurately. For example, under current PTB tokenization policy, the shorthand notation 'd/c' (for the verb 'discontinue') would be annotated as three tokens: d/AFX//HYPH c/VB; however, we decided to annotate the abbreviation as one token (d/c/VB) to better align with the full form of the verb.

Treebanking the clinical narrative entails several phases of automatic preprocessing and manual correction of each layer of output. First, all formatting metadata are stripped from the original source files. Then, the data are segmented into individual sentences. These sentence units are fed through an automatic tokenizer and then a POS tagger. Manual correction of segmentation, tokenization, and POS tagging takes place before the

[i] Contact Guergana.Savova@childrens.harvard.edu to initiate the data use agreement.



Example text from clinical note:

The patient underwent a radical tonsillectomy (with additional right neck dissection) for metastatic squamous cell carcinoma. He returns with a recent history of active bleeding from his oropharynx.

Example Treebank annotations:

```
((S (NP-SBJ (DT The)
            (NN patient))
    (VP (VBD underwent)
        (NP (NP (DT a)
                (JJ radical)
                (NN tonsillectomy))
            (-LRB- ()
            (PP (IN with)
                (NP (JJ additional)
                    (NML (JJ right)
                         (NN neck))
                    (NN dissection)))
            (-RRB- ))
            (PP (IN for)
                (NP (JJ metastatic)
                    (NML (JJ squamous)
                         (NN cell))
                    (NN carcinoma)))))
    (. .)) )

((S (NP-SBJ (PRP He))
    (VP (VBZ returns)
        (PP (IN with)
            (NP (NP (DT a)
                    (JJ recent)
                    (NN history))
                (PP (IN of)
                    (NP (NP (JJ active)
                            (NN bleeding))
                        (PP (IN from)
                            (NP (PRP$ his)
                                (NN oropharynx)))))))))
    (. .)) )
```

Example PropBank annotations:

REL: undergo.01
ARG1: the patient
ARG2: radical tonsillectomy (with additional right neck dissection)

REL: return.01
ARG1: he
ARGM-ADV: with a recent history of active bleeding from his oropharynx

REL: bleeding.01
ARG1: oropharynx
ARGM-ADJ: active

Example UMLS annotations:

Entities
[patient]: Person
[radical tonsillectomy (with additional right neck dissection)]: Procedure
[radical tonsillectomy]: Procedure
[additional right neck dissection]: Procedure
[right neck]: Anatomy
[metastatic squamous cell carcinoma]: Disorder
[active bleeding from his oropharynx]: Disorder
[active bleeding]: Disorder
[oropharynx]: Anatomy

Figure 1 Example text from a clinical note with Treebank, PropBank and UMLS annotations.

data are automatically syntactically parsed with the Bikel parser.[25] The constituency parse trees are manually corrected and empty categories and function tags are added. Empty categories include elided arguments such as dropped subjects (*PRO*) or moved arguments such as passive traces (*T*). Function tags include argument labels such as –SBJ (subject) and –PRD (predicate), and adjunct descriptors such as –LOC (locative) and –TMP (temporal). During syntactic correction any lingering segmentation, tokenization, or POS tagging errors are also corrected. All files receive a second pass of tree correction. After all files have been through the above automatic and manual processes, quality control checks and validation scripts are run. Completed data provide gold-standard constituent structure trees with function tags and traces.

Due to the resource-intensive nature of Treebanking, only a small set (about 8% of the total completed data) was double-annotated to calculate inter-annotator agreement (IAA). The startup cost for this annotation project was quite high ($70 000) since guidelines had to be adapted to the fragmentary data and annotators had to be trained in both syntactic structure and medical terminology, giving an overall estimated cost for Treebanking these data of close to $100 000.

In our previous work,[26] we showed results on training the OpenNLP constituency parser on a corpus combining general domain data and the MiPACQ data described in this manuscript. The parser achieved a labeled F_1 score of 0.81 on a corpus consisting of CN and pathology notes when tested on held-out data of clinical and pathology notes. This result is lower than the general domain state of the art but difficult to contextualize due to the lack of other reported work in the clinical domain. However, it is similar to the performance of the dependency parser described in this manuscript.

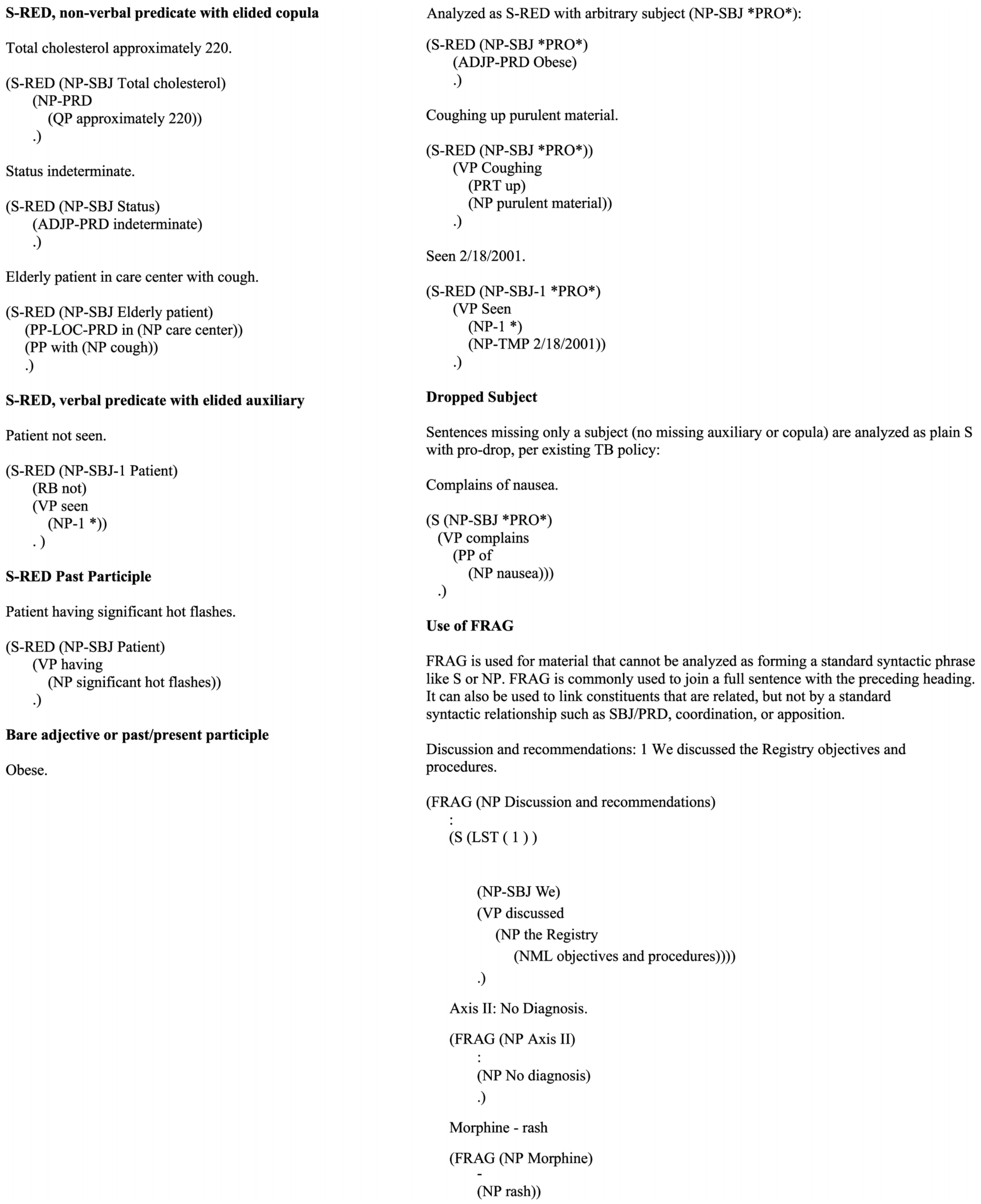
PropBank annotations

The goal of this layer of annotations is to mark the predicate-argument structure of sentences[4] (see figure 1 for an example). PropBank annotation consists of two main stages: (1) creation of frame files for predicates (verbs and nominative predicates) occurring in the data; and (2) annotation of the data using the argument structures outlined in the frame files.

Each frame file may contain one or more framesets, corresponding to coarse-grained senses of the predicate lemma. For example, the frame file for 'indicate' contains a frameset for 'show' and a frameset for 'recommend a course of action,' each of these senses having different arguments (the second sense being especially common in medical documents). Each frameset outlines the semantic roles that are possible or commonly used with a given predicate, and numbers these arguments consistently across predicates. See table 1 for details on arguments and matching roles.

Linguists create the framesets as they occur in the data. The frame creators draw assistance from lexical resources including VerbNet[27] and FrameNet,[28] as well as from the framesets of analogous predicates. The framesets contain numbered argument roles corresponding to those arguments most closely related to the predicate, but the annotators also use broader annotation labels, ArgMs, which include supplementary arguments, such as manner (ARG-MNR), location (ARG-LOC), and temporal information (ARG-TMP). See figure 1 for a running example.

Data that have been annotated for Treebank syntactic structure and have had frame files created are passed on to the PropBank annotators for double-blind annotation. The annotators determine which sense of a predicate is being used, select the corresponding frame file, and label the occurring arguments as outlined in the frame file. This task relies on the syntactic annotations done in Treebanking, which determine the span of constituents such as verb phrases, which then set the boundaries for PropBank annotation. Once a set of data has been double-annotated, it is passed on to an adjudicator, who resolves any disagreements between the two primary annotators to create the gold standard. Even with the double annotation, adjudication, and new frame file creation, the cost of PropBanking these data is less than half the cost of Treebanking, or approximately $40 000, with less than half of that for the ramp-up cost.



S-RED, non-verbal predicate with elided copula

Total cholesterol approximately 220.

```
(S-RED (NP-SBJ Total cholesterol)
     (NP-PRD
        (QP approximately 220))
     .)
```

Status indeterminate.

```
(S-RED (NP-SBJ Status)
     (ADJP-PRD indeterminate)
     .)
```

Elderly patient in care center with cough.

```
(S-RED (NP-SBJ Elderly patient)
    (PP-LOC-PRD in (NP care center))
    (PP with (NP cough))
    .)
```

S-RED, verbal predicate with elided auxiliary

Patient not seen.

```
(S-RED (NP-SBJ-1 Patient)
     (RB not)
     (VP seen
        (NP-1 *))
     . )
```

S-RED Past Participle

Patient having significant hot flashes.

```
(S-RED (NP-SBJ Patient)
     (VP having
        (NP significant hot flashes))
     .)
```

Bare adjective or past/present participle

Obese.

Analyzed as S-RED with arbitrary subject (NP-SBJ *PRO*):

```
(S-RED (NP-SBJ *PRO*)
     (ADJP-PRD Obese)
     .)
```

Coughing up purulent material.

```
(S-RED (NP-SBJ *PRO*))
     (VP Coughing
        (PRT up)
        (NP purulent material))
     .)
```

Seen 2/18/2001.

```
(S-RED (NP-SBJ-1 *PRO*)
     (VP Seen
        (NP-1 *)
        (NP-TMP 2/18/2001))
     .)
```

Dropped Subject

Sentences missing only a subject (no missing auxiliary or copula) are analyzed as plain S with pro-drop, per existing TB policy:

Complains of nausea.

```
(S (NP-SBJ *PRO*)
  (VP complains
     (PP of
        (NP nausea)))
  .)
```

Use of FRAG

FRAG is used for material that cannot be analyzed as forming a standard syntactic phrase like S or NP. FRAG is commonly used to join a full sentence with the preceding heading. It can also be used to link constituents that are related, but not by a standard syntactic relationship such as SBJ/PRD, coordination, or apposition.

Discussion and recommendations: 1 We discussed the Registry objectives and procedures.

```
(FRAG (NP Discussion and recommendations)
    :
    (S (LST ( 1 ) )


         (NP-SBJ We)
         (VP discussed
            (NP the Registry
               (NML objectives and procedures))))
         .)
```

Axis II: No Diagnosis.

```
(FRAG (NP Axis II)
     :
     (NP No diagnosis)
     .)
```

Morphine - rash

```
(FRAG (NP Morphine)
     -
     (NP rash))
```

Figure 2 Example of clinical Treebank guideline changes.

UMLS entities

We adopted the UMLS semantic network for semantic annotation of NEs.[13] We chose to limit our use of semantic network entity types to mostly semantic groups.[13] By making this choice, our annotators would not have to differentiate, for instance, between a 'Cell or Molecular Dysfunction' and a 'Neoplastic Process,' instead using 'Disorder.' In addition to reducing errors due to lack of specific domain knowledge, this resulted in more tokens per entity type, increasing statistical power for classification. Also, these broad semantic groups are helpful for normalization against community-adopted conventions such as the Clinical Element Model[29] whose core semantic types are

Table 1 Sample sentence and PropBank frame and roles

Argument role	Predicate argument	Example: frame for 'decrease'	Example: 'Dr Brown decreased the dosage of Mr Green's medications by 20 mg, from 50 mg to 30 mg, in order to reduce his nausea'
Arg0	Agent	Causer of decline, agent	Dr Brown
Arg1	Patient	Thing decreasing	The dosage of Mr Green's medication
Arg2	Instrument, benefactive, or attribute	Amount decreased by; extent or manner	By 20 mg
Arg3	Starting point, benefactive, or attribute	Start point	From 50 mg
Arg4	Ending point	End point	To 30 mg
ArgM	Modifier		ArgM-PRP (purpose): in order to reduce his nausea

Disorders, Sign or Symptoms, Procedures, Medications, and Labs. The Sign or Symptom semantic type was annotated as a semantic category independent of the Disorders semantic group because many applications such as phenotype extraction and clinical question answering require differentiations between Disorders and Sign/Symptom. Each UMLS entity has two attribute slots: (1) Negation, which accepts *true* and *false* (default) values; and (2) Status, which accepts *none* (default), *Possible*, *HistoryOf*, and *FamilyHistoryOf* values.

To the set of UMLS semantic categories we added the Person category to align the annotations with the definitions in the general domain. We felt that the UMLS semantic group of Living Beings is too broad, while the UMLS semantic types of Human, Patient or Disabled Group, Family Group, Age Group, Population Group, and Professional or Occupational Group presented definitional ambiguities.

The corpus was pre-annotated for UMLS entities with cTAKES.[14] Seventy-four percent of the tokens in the MiPACQ corpus were annotated in parallel by two annotators. The remaining 26% were single-annotated. Double-annotated data were adjudicated by a medical expert, creating the gold standard.

Example UMLS entities are shown in figure 1. The cost of UMLS annotation is somewhat higher than that of PropBanking, mainly because of the time involved in consulting the UMLS ontology, at about $50 000–$60 000 for these data with about a third of it for the ramp-up cost.

IAA and evaluation metrics

IAA is reported. The annotations of one annotator were used as the gold standard against which to calculate the precision, recall, and F_1 measure of the second annotator.[30]

The agreement figures on the Treebank data were calculated using EvalB,[31] the most commonly used software for comparing bracketed trees, to compute IAA as an F_1 measure.[30] When comparing trees, constituents are said to match if they share the same node label and span; punctuation placement, function tags, trace and gap indices, and empty categories are ignored.

The agreement on PropBank data was computed using exact pairwise matches for numbered arguments and ArgMs. In the calculation of PropBank agreement, one 'annotation' consists of at least two different constituents linked by a particular roleset. Two annotations were counted as an 'exact' match if their constituent boundaries and roles (numbered arguments, Arg0–5, and adjunct/function or ArgM types) matched. Two annotations were counted as a 'core-arg' match if their constituent boundaries matched and they had the same role number or were both ArgM types (in this type of match, the difference between ArgM-MNR and ArgM-TMP is ignored; as long as both annotators have used a type of ArgM, a match is counted). A 'constituent' match was counted if the annotators marked the same constituent.

To compute UMLS IAA, we aligned the entities of two annotators using a dynamic programming alignment algorithm.[32 26] IAA was computed as the F_1 measure between annotators.[30] We report two types of IAA that correspond to two different approaches to aligning the spans of the entity mentions. The first requires that the boundaries of the compared entity mention spans match exactly, while the second allows partial matching.

RESULTS

Corpus characteristics

The final corpus consists of 13 091 Treebanked sentences. For the PropBanking layer of annotations, the MiPACQ data included usages of 1772 distinct predicate lemmas, of which 1006 had existing frame files prior to the beginning of the project. Of the 766 newly created frame files, only 74 were verbs (the rest being predicating nouns). For example, new frames were created for *adrenalectomize*, *clot*, *disinfest*, *excrete*, *herniated*, *protrusion*, *ossification*, *palpitation*, and *laceration*, which are specific to the clinical domain. The PropBank database can be accessed at the PropBank website.[33]

The majority of the annotations of numbered arguments were to Arg0s (48.47%), with decreasing numbers of annotations for higher-numbered arguments (table 2). This distribution is fairly representative of PropBank annotations in general. The distribution of NE annotations over the Person and UMLS semantic categories total 28 539 spread over 15 UMLS semantic groups, one UMLS semantic type, and the Person semantic category (table 2).

IAA metrics aim at estimating the quality of the gold standard and are often considered a high bar for the expected system performance, although they are not a strict upper-bound. The MiPACQ corpus IAA results (table 3) are strong, implying computational learnability. POS tagging IAA is typically significantly higher than parse IAA.

Development and evaluation of NLP components

We built several NLP components using the MiPACQ corpus—POS tagger, constituency parser, dependency parser, and SRL[26 34–36]—which were released as a part of cTAKES.[14 37] To build statistical models for POS tagging, dependency parsing (DEP), and dependency-based SRL, we divided the corpus into training, development, and evaluation sets (85%, 5%, and 10%, respectively).

The first step in building the dependency parser involves the conversion of the constituent trees in the Treebank into dependency trees. The constituent-to-dependency conversion was done by using the Clear dependency convertor,[38] although a major

Table 2 Frequency of annotations

Annotation type	Raw annotation counts	In %
PropBank argument: Arg0	9647	48.47
PropBank argument: Arg1	2901	14.58
PropBank argument: Arg2	146	0.73
PropBank argument: Arg3	109	0.55
PropBank argument: Arg4	0	0.00
PropBank argument: ArgM	7098	35.67
PropBank argument: Total	*19901*	*100.00*
UMLS semantic group: Procedures	4483	15.71
UMLS semantic group: Disorders	4208	14.74
UMLS semantic group: Concepts and Ideas	4308	15.10
UMLS semantic group: Anatomy	3652	12.80
UMLS semantic type: Sign or Symptom	3556	12.46
UMLS semantic group: Chemicals and Drugs	2137	7.49
UMLS semantic group: Physiology	1669	5.85
UMLS semantic group: Activities and Behaviors	990	3.47
UMLS semantic group: Phenomena	847	2.97
UMLS semantic group: Devices	282	0.99
UMLS semantic group: Living Beings	120	0.42
UMLS semantic group: Objects	103	0.36
UMLS semantic group: Geographic Areas	84	0.29
UMLS semantic group: Organizations	60	0.21
UMLS semantic group: Occupations	24	0.08
UMLS semantic group: Genes and Molecular Sequences	1	0.00
Non-UMLS semantic category: Person	2015	7.06
UMLS and non-UMLS semantic annotations: Total	*28539*	*100.00*

UMLS, Unified Medical Language System.

Table 3 Inter-annotator agreement results (F_1 measure)

	Average IAA
Treebank	0.926
PropBank, exact	0.891
PropBank, Core-arg	0.917
PropBank, Constituent	0.931
UMLS, exact	0.697
UMLS, partial	0.750

IAA, inter-annotator agreement; UMLS, Unified Medical Language System.

change was made in the dependency format. The Clear dependency convertor generates the CoNLL dependencies[39] used for the CoNLL'08-09 shared tasks.[40 41] However, we recently discovered that several NLP components have started using the Stanford dependencies[42]; thus, we adapted our Clear dependency conversion to the Stanford dependencies, with an important modification. The original Stanford dependency convertor does not produce long-distance dependencies. Our approach produces the same long-distance dependencies provided by the CoNLL dependency conversion while using the more fine-grained Stanford dependency labels.[ii]

To build a POS tagging model, we used Apache OpenNLP.[43] The OpenNLP POS tagger uses maximum entropy for learning and a simple one-pass, left-to-right algorithm for tagging. To build a DEP model, we used the Clear dependencies described above with Liblinear for learning[38] and a transition-based DEP algorithm for parsing.[34 35] The parsing algorithm used in the Clear dependency parser can generate both local dependencies (projective dependencies) and long-distance dependencies (non-projective dependencies) without going through extra steps of pre- or post-processing. Non-projective DEP can generally be done in quadratic time. However, our experiments showed that in practice the Clear dependency parser gives a linear-time parsing speed for non-projective parsing (taking about 2–3 ms per sentence) while showing comparable accuracies against other state-of-the-art dependency parsers.[34]

To build a dependency-based SRL model, we used the Clear SLR.[36] Unlike constituent-based SRL where semantic roles are assigned to phrases (or clauses), semantic roles are assigned to the headwords of phrases in dependency-based SRL. This may lead to a concern about getting the actual semantic spans back, but a previous study has shown that it is possible to recover the original spans from the headwords with minimal loss, using a certain type of dependency structure.[39] The Clear SLR also uses Liblinear for learning and a transition-based algorithm for labeling, which compares each identified predicate to all other word-tokens in a sentence for finding its semantic arguments.

Table 4, part A provides details of the number of tokens and sentences in each corpus we used to train a different model. We used two different versions of the WSJ Treebank, the standard OntoNotes version and a smaller subset equivalent in size to our MiPACQ training corpus. We also evaluated the performance when the MiPACQ corpus was combined with both the small and the large WSJ corpora.

Table 4, part B provides similar details of our test sets. These include two different MiPACQ sets, since we separated out the very formulaic pathology notes (MiPACQ-PA) so they would not artificially increase our performance on standard CN (MiPACQ-CN). To test portability to other genres and note styles, we also tested on (1) radiology notes from the Strategic

Table 4 The distribution of the training data across the different corpora

Part A

	WSJ (901 K)	WSJ (147 K)	MiPACQ (147 K)	WSJ +MiPACQ (147 K +147 K)	WSJ +MiPACQ (901 K +147 K)
# Of sentences	37015	6006	11435	17441	43021
# Of word-tokens	901673	147710	147698	295408	1049383
# Of verb-predicates	96159	15695	16776	32471	111854

Part B

	MiPACQ-CN	MiPACQ-PA	SHARP	THYME
Genre	Colon cancer	Pathology	Radiology	Colon cancer
# Of sentences	893	203	9070	9107
# Of word-tokens	10865	2701	119912	102745
# Of verb-predicates	1355	145	8573	8866

CN, clinical notes; PA, pathology notes; WSJ, Wall Street Journal.

[ii] A long-distance dependency is a dependency relation between a pair of word-tokens that are not within the same domain of locality.

Health IT Advanced Research Project: Area 4 project (SHARP[44]), and (2) colon cancer clinical and pathology notes from Temporal Histories of Your Medical Events, an NIH Project on Temporal Reasoning (THYME[45]).

Table 5 shows results for POS tagging (POS), dependency parsing (DEP), and dependency-based semantic role labeling (SRL), respectively, using the training and evaluation sets described above. ACC shows accuracies (in %). UAS and LAS show unlabeled attachment scores and labeled attachment scores (in %). $AI\!-\!F_1$ and $AI\!+\!AC\!-\!F_1$ show F_1 scores of argument identification and both identification and classification (in %), respectively.

The results on the formulaic MiPACQ pathology notes, as expected, were very high. POS tagging is 98.67%, LAS for DEP is 89.23%, and the F_1 score for argument identification and classification for SRL is 90.87%. It is unusual to encounter new phenomena in these notes that have not already been seen in the training data.

DISCUSSION

The layered syntactic and semantic annotations of the MiPACQ clinical corpus present the community with a new comprehensive annotated resource for further research and development in clinical NLP. One of the primary ways in which these annotations will be used is in the creation of NLP components as described in the 'Development and evaluation of NLP components' section, for use in tasks such as information extraction, question answering, and summarization to name a few. The 766 biomedical-related PropBank frame files created for this project are available on-line and add to the comprehensiveness of existing frame files. As expected, the existence of domain-specific annotations improves the accuracy for all of the NLP components.

The agreement on Treebank and PropBank annotations is similar to that reported in the general domain. The agreement on the UMLS annotations is similar to results reported previously.[17] Rich semantic annotations of the type described in this paper are the building blocks to more complex tasks, such as the discovery of implicit arguments and inferencing.[46] Systems for analyzing these complicated phenomena will require a large number of resources to ensure accuracy and efficacy. The annotations completed in this project provide complementary information to other annotated resources, such as Informatics for Integrating Biology at the Bedside (i2b2)[47] and Ontology Development and Information Extraction,[48 49] helping to create a more complete set of resources.

Table 5 Evaluation on the MiPACQ corpus/SHARP corpus/THYME corpus

	Evaluation metric	WSJ (901 K)	WSJ (147 K)	MiPACQ (147 K)	WSJ +MiPACQ (147 K +147 K)	WSJ +MiPACQ (901 K +147 K)
POS	ACC	88.62	87.79	94.28	94.39	94.11
		81.71	81.38	90.13	89.17	87.59
		83.07	82.32	92.12	92.00	90.84
DEP	UAS	78.34	75.59	85.72	85.30	85.40
		67.34	65.01	74.93	74.70	73.89
		65.58	62.11	73.21	73.56	73.95
	LAS	74.37	70.40	83.63	83.23	83.31
		62.63	59.09	72.19	71.80	70.35
		60.23	56.33	70.26	70.76	70.96
SRL	$AI\!-\!F_1$	76.98	74.57	86.58	87.31	88.17
		74.29	71.57	80.86	82.86	82.66
		74.16	72.17	86.20	86.69	86.29
	$AI\!+\!AC\!-\!F_1$	67.63	63.44	77.72	79.35	79.91
		62.16	57.03	69.43	71.64	72.00
		63.28	58.32	76.69	77.46	78.38

DEP, dependency parsing; LAS, labeled attachment scores; POS, part-of-speech; SLR, semantic role labeler; UAS, unlabeled attachment scores; WSJ, Wall Street Journal.

The results that we report here on critical NLP components—POS tagger, dependency parser, and SRL—can be used as baselines in the clinical NLP community, especially given the fact that the dependency parser and the SRL components are among the first of their kind in the clinical domain. Semantic role labeling is still a relatively new representation in the NLP community, but it is beginning to contribute to significant improvements in question answering and coreference.[50 51] Within the domain of biomedicine, one immediate application is the discovery of the subject of a clinical event, be it the patient, a family member, donor, or other. cTAKES implements such a module which directly consumes the output of the semantic role labeling described in this manuscript. As our NLP results improve to the point where they will better support these types of applications, we expect to see increased interest in semantic role labeling. Our current efforts on which we will be reporting separately are focusing on the higher level components such as UMLS named entity recognition (NER).

For POS tagging, DEP, and dependency-based SRL, all models trained on the MiPACQ corpus show significant improvements over those trained on the WSJ corpus (McNemar, $p<0.0001$), although the MiPACQ models are trained on far fewer data. This implies that having even a small amount of in-domain annotation can enhance the quality of these NLP components for this domain. We expect these data to provide some improvement in all clinical narrative data, but the more divergent the data are from our corpus, the less improvement there will be. This approach to NLP still requires additional domain-specific annotations for the highest possible performance, and the clinical arena has a multitude of domains. Various domain-adaptation techniques can be applied to improve the performance, which we will explore in the future. Notice that the MiPACQ models show higher accuracies for MiPACQ-PA than MiPACQ-CN in all three tasks. From our error analysis, we found that sentences in MiPACQ-PA were very uniform among themselves , so not many training data are needed.

The MiPACQ models also showed a significant improvement over the WSJ models for our other test sets, SHARP and THYME. The performance is almost 10% lower than for MiPACQ. However, notice that the WSJ performance is also 10% lower. These test sets have more sentence fragments and novel vocabulary than the MiPACQ test set, and are correspondingly more challenging. The advantage of being able to develop annotations in close collaboration with the Penn Treebankers is reflected in the consistent parsing results across genres.

One of the most significant limitations that has been uncovered by this project is the need for a widely agreed-upon annotation schema for the clinical NLP community. While the UMLS is generally accepted as the go-to schema for semantic annotation of biomedical information, this project has highlighted shortcomings. As the UMLS schema was not originally designed for annotation use, the definitions of some of the semantic types are closely overlapping. For example, there did not seem to be a clear semantic group or category unambiguously encompassing Person mentions. This necessitated the usage of the non-UMLS Person semantic category. In addition, the sheer size of the UMLS schema increases the complexity of the annotation task and slows annotation, while only a small proportion of the annotation types present are used. Because of this level of complexity, we chose to annotate predominantly at the level of the UMLS semantic groups. In our project motivated by user requirements we chose to refine the annotations by using the UMLS semantic type Sign or Symptom by itself independently of the Disorder semantic group. The schema adaptations discussed in the present article must be viewed not as defining a finished schema, but as the initial significant steps required for comprehensive annotation. As the UMLS semantic hierarchy allows adaptations, standardization across different groups and projects would enhance collaboration efforts.

Although the CN used in this project provide a large variety of syntactic and semantic information, the performance of any biomedical NLP system would only be improved by the annotation of additional types of data, such as discharge summaries, emergency department notes, and call transcripts. Under SHARP4[44] we are annotating a 500 K word corpus consisting of a representative set of CN from two institutions. Our starting point is the layered annotations, guidelines, and schemas described here. To that, we are adding coreference and temporal relations.[52] For each layer, our future efforts will involve developing generalizable principled algorithms. We will also explore active learning methods to decrease the annotation cost without sacrificing annotation quality.[53 54]

CONCLUSION

In this paper, we have described a foundational step towards building a manually annotated lexical resource for the clinical NLP community using three layers of annotation: constituency syntax (Treebank), predicate-argument structure (PropBank), and clinical entities (UMLS). By using this multi-layered approach, we have created the first syntactically and semantically annotated clinical corpus, and this has major implications for clinical NLP in general. In addition to enabling the biomedical field to take advantage of previously developed NLP tools, this project has shown that syntactic and semantic information can be effectively annotated in clinical corpora, and that a reasonable level of inter-annotator agreement and NLP component performance can be achieved for all of these annotation layers.

Acknowledgements We would like to thank our domain experts, Dr William F Styler III and Donna Ihrke. James Masanz from the Mayo Clinic provided outstanding technical support. We are very grateful for the help on UMLS annotations Drs Thomas Rindflesch and Marcelo Fiszman of the NLM provided along the way.

Contributors All authors of this manuscript contributed to (1) conception and design, acquisition of data or analysis and interpretation of data, (2) drafting the article or revising it critically for important intellectual content, and (3) final approval of the version published.

Funding The project described was supported by award numbers NLM RC1LM010608 (MiPACQ), NLM 10090 (THYME), and 90TR0002 (SHARP). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NLM/NIH/ONC.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The corpus can be obtained through a data use agreement as indicated in the manuscript. The software is open source and URLs are provided for downloading.

REFERENCES

1 Chapman W, Nadkarni P, Hirschman L, *et al*. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *JAMA* 2011;18:540–3.

2 Marcus M, Santorini B, Marcinkiewicz M. Building a large annotated corpus of English: The Penn Treebank. *Comput Linguist* 1994;19:313–30.

3 Prasad R, Lee A, Dinesh N, *et al*. Penn discourse Treebank version 2.0. *Linguistic Data Consortium* 2008, February.

4 Palmer M, Gildea D, Kingsbury P. The proposition bank: a corpus annotated with semantic roles. *Comput Linguist* 2005;31:71–106.

5 Santorini B. *Part-of-speech tagging guidelines for the Penn Treebank Project*. MS-CIS-90-47, Technical report. Department of Computer and Information Science, University of Pennsylvania; 1990.

6 Bies A, Ferguson M, Katz K, *et al*. *Bracketing guidelines for Treebank II Style Penn Treebank project*. ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz (accessed 15 Aug 2012).

7 Taylor A. *Bracketing Switchboard: an addendum to the Treebank II bracketing guidelines*. http://www.cis.upenn.edu/~bies/manuals/prsguid2.pdf (accessed 15 Aug 2012).

8 Warner C, Bies A, Brisson C, *et al*. *Addendum to the Penn Treebank II style bracketing guidelines: BioMedical treebank annotation*. http://papers.ldc.upenn.edu/Treebank_BioMedical_Addendum/TBguidelines-addendum.pdf (accessed 15 Aug 2012).

9 Taylor A. *Treebank 2a guidelines*. http://www-users.york.ac.uk/~lang22/TB2a_Guidelines.htm (accessed 15 Aug 2012).

10 Mott J, Warner C, Bies A, *et al*. *Supplementary guidelines for English translation Treebank 2.0*. http://projects.ldc.upenn.edu/gale/task_specifications/ettb_guidelines.pdf (accessed 15 Aug 2012).

11 Taylor A. *Reconciliation of differences between Onto/WSJ and EXTB*. http://www.seas.upenn.edu/~jmott/Update-2010-ETTB-Onto-reconciliation.pdf (accessed 15 Aug 2012).

12 *Unified Medical Language System (UMLS)*. http://www.nlm.nih.gov/research/umls/ (accessed 15 Aug 2012).

13 Bodenreider O, McCray A. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003;36:414–32.

14 *Clinical Text Analysis and Knowledge Extraction System (cTAKES)*. http://incubator.apache.org/ctakes/ (accessed 15 Aug 2012).

15 Nielsen R, Masanz J, Ogren P, *et al*. An architecture for complex clinical question answering. *1st Annual ACM International Conference on Health Informatics (IHI 2010)*. Washington, DC, 2010:395–9.

16 Cairns B, Nielsen R, Masanz J, *et al*. The MiPACQ clinical question answering system. *Proceedings of the American Medical Informatics Association annual symposium*. Washington, DC, 2011:171–80.

17 Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. *Language Resources and Evaluation Conference (LREC)*. Marakesh, Morrocco, 2008:3143–50. http://www.lrec-conf.org/proceedings/lrec2008/ (accessed 10 Jan 2013).

18 Roberts A, Gaizauskas R, Hepple M, *et al*. Building a semantically annotated corpus of clinical text. J Biomed Inform, 2009;42:950–66.

19 Sauri R, Littman J, Knippen B, *et al*. *TimeML annotation guidelines*. 2006. http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf (accessed 15 Aug 2012).

20 Uzuner O, South B, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;18:552–6.

21 Vincze V, Szarvas G, Farkas R, *et al*. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics* 2008;9(Suppl 11):S9.

22 *Treebanking annotation guidelines*. http://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf (accessed 10 Jan 2013).

23 *PropBanking annotation guidelines*. http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf (accessed 10 Jan 2013).

24 *UMLS annotation guidelines*. http://clear.colorado.edu/compsem/documents/umls_guidelines.pdf (accessed 10 Jan 2013).

25 Bikel D. *Multilingual statistical parsing engine*. http://www.cis.upenn.edu/~dbikel/software.html#stat-parser (accessed 15 Aug 2012).

26 Zheng J, Chapman W, Miller T, *et al*. A system for coreference resolution for the clinical narrative. *J Am Med Inform Assoc* doi:10.1136/amiajnl-2011-000599.

27 Kipper-Schuler K. VerbNet: a broad coverage, comprehensive verb lexicon. Philadelphia, PA: Department of Computer Science, University of Pennsylvania, 2005.

28 Baker C, Fillmore C, Lowe J. The Berkley FrameNet Project. *Conference on Computational Linguistics (COLING/ACL-98)*. Montreal, Canada, 1998:86–90.

29 *Clinical Element Models (CEMs)*. http://www.clinicalelement.com (accessed 15 Aug 2012).

30 Hripcsak G, Rothschild A. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005;12:296–8.

31 Sekine S, Collins M. *EvalB*. http://nlp.cs.nyu.edu/evalb/ (accessed 15 Aug 2012).

32 Cormen T, Leiserson C, Rivest R, *et al*. *Introduction to algorithms*. Cambridge, MA: Massachusetts Institute of Technology, 2009.

33 http://verbs.colorado.edu/propbank/framesets-english/ (accessed 10 Jan 2013).

34 Choi J, Palmer M. Getting the most out of transition-based dependency parsing. *46th Annual meeting of the Association for Computational Linguistics and Human Language Technologies*. Portland, OR, 2011:687–92.

35 Choi J, Nicolov N. K-best, transition-based dependency parsing using robust minimization and automatic feature reduction. Collections of Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, 2010:288–302.

36 Choi JD, Palmer M. *Transition-based semantic role labeling using predicate argument clustering*. Portland, OR: Association of Computational Linguistics workshop on Relational Models of Semantics, 2011:37–45.

37 Savova G, Masanz J, Ogren P, *et al*. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.

38 Choi JD, Palmer M. Robust constituent-to-dependency conversion for English. *9th International Workshop on TreebankTreebanks and Linguistic Theories*. Tartu, Estonia, 2010:55–66.

39 Richard J, Nugues P. Extended constituent-to-dependency conversion for English. *The 16th Nordic Conference of Computational Linguistics*, 2007.

40 Surdeanu M, Johansson R, Meyers A, *et al*. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. *12th Conference on Computational Natural Language Learning (CoNLL): Shared Task*, 2008:59–177.

41 Hajic J, Ciaramita M, Johansson R, *et al*. The CoNLL-2009 Shared Task on Syntactic and semantic dependencies in multiple languages. *13th Conference on Computational Natural Language Learning (CoNLL): Shared Task*. 2009:1–18.

42 de Marneffe M, Manning C. The Stanford typed dependencies representation. COLING workshop on Cross-Framework and Cross-Domain Parser Evaluation. 2008.

43 Apache OpenNLP. http://opennlp.apache.org (accessed 15 Aug 2012).

44 Strategic Health IT Advanced Research Projects: Area 4 (SHARPn): Secondary data use and normalization. http://sharpn.org (accessed 31 Jul 2011).

45 Temporal Histories of Your Medical Events, THYME. https://clear.colorado.edu/TemporalWiki/index.php/Main_Page (accessed 15 Aug 2012).

46 Gerber M, Chai J. Beyond NomBank: a study of implicit arguments for nominal predicates. *48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, 2010:1583–92.

47 http://www.i2b2.org (accessed 10 Jan 2013).

48 Savova G, Chapman W, Zheng J, *et al*. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc* 2011;18:459–65.

49 Chapman W, Savova G, Zheng J, *et al*. Anaphoric reference in clinical reports: characteristics of an annotated corpus. *J Biomed Inform* 2012;45:507–21.

50 Lee H, Recasens M, Chang A, *et al*. Joint entity and event coreference resolution across documents. *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, 2012.

51 Surdeanu M, Ciaramita M, Zaragoza H. Learning to rank answers to non-factoid questions from web collections. *Comput Linguist* 2011;37:351–83.

52 Savova G, Bethard S, Styler W, *et al*. *Towards temporal relation discovery from the clinical narrative*. San Francisco, CA: American Medical Informatics Association Annual Symposium, 2009:568–72.

53 Settles B. Active learning literature survey. Technical Report 1648, University of Wisconsin—Madison, Computer Sciences Technical Report 1648, 2010. Computer Sciences Technical Report 1648.

54 Miller T, Dligach D, Savova G. Active learning for Coreference Resolution in the Biomedical Domain. *BioNLP workshop at the Conference of the North American Association of Computational Linguistics (NAACL)*, 2012.