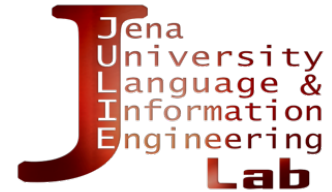




**FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA**



Annotationsleitlinien für deutschsprachige Medizintexte.

Teil 2: Annotation von personenidentifizierenden (PHI-)Attributen

Tobias Kolditz, Christina Lohr, Luise Modersohn & Udo Hahn

Jena University Language & Information Engineering Lab (JULIE Lab)

Friedrich-Schiller-Universität Jena

JULIE Lab – SMITH Technischer Report 2

<https://doi.org/10.5281/zenodo.7707882>

Dezember 2022

Inhaltsverzeichnis

1	Vorwort	2
2	Einleitung	3
3	Daten	3
4	Annotationsleitlinien	4
4.1	Allgemeine Hinweise	4
4.2	Typensystem für PHI-Annotationen	4
4.2.1	Person	4
4.2.2	Date	4
4.2.3	Age	4
4.2.4	Location	5
4.2.5	MedicalUnit	5
4.2.6	ID	5
4.2.7	Contact	5
4.3	Other	5
5	Das Annotationswerkzeug BRAT	5
5.1	Hinweise zur Konfiguration von BRAT	6
5.2	Hinweise zur Nutzung von BRAT	6
5.2.1	Hinweise zum Annotationswerkzeug BRAT	6
5.2.2	Inhaltliche Hinweise zur Annotation	7
5.2.3	Was nicht zu annotieren ist	8
A	Appendix	10

1 Vorwort

Die hier dokumentierten Annotationsleitlinien für deutschsprachige Medizintexte sind im Rahmen des SMITH-Projekts¹ im Zeitraum zwischen 2018 bis 2022 entstanden.

Der vorliegende Teil 2 hat die Kennzeichnung von potenziell personenidentifizierenden Attributen in klinischen Dokumenten zum Gegenstand. Er spiegelt die auch gesetzlich verankerten Restriktionen wider, die die Distribution klinischer Daten (hier in Texten repräsentiert) jenseits der klinischen Institutionen, in denen diese Daten (Texte) entstanden sind, einschränken. Um eine darüberhinausgehende Distribution überhaupt zu ermöglichen, müssen personenidentifizierende Attribute (etwa Namen, Adressen, Datumsangaben) *de-identifiziert* werden. Wir modellieren diese Attribute typbezogen, um ein hohes Maß an Flexibilität bei der nachfolgenden automatischen Erkennung von personenidentifizierenden Entitäten (*named entities*) und möglichen Transformation (etwa in Form der Pseudonymisierung) sensibler Textpassagen zu ermöglichen.

Anders als in den USA, wo eine Liste von 18 solcher identifikationsrelevanten Attribute (sog. *Protected Health Information* – PHI) im *Health Insurance Portability and Accountability Act* (HIPAA)² gesetzlich fixiert wurde, liegen in der Bundesrepublik Deutschland keine vergleichbaren Listen vor. Mit der vorliegenden Annotationsleitlinie wird ein solcher Vorschlag gemacht — er lehnt sich an die US-amerikanischen PHI-Kriterien an, adaptiert sie jedoch gleichzeitig an bundesdeutsche Gegebenheiten.

¹Das SMITH-Projekt (<https://www.smith.care>) wurde im Rahmen der Medizininformatik-Initiative (MII – <https://www.medizininformatik-initiative.de>) vom Bundesministerium für Bildung und Forschung (BMBF) gefördert (Fördernummer: 01ZZ1803G).

²<https://www.cdc.gov/phlp/publications/topic/hipaa.html>

2 Einleitung

Einer der Garanten für die aktuellen Erfolge bei der automatischen Verarbeitung natürlicher Sprache ist die Verfügbarkeit von (möglichst großvolumigen) Sprachdaten. Sind solche Datensammlungen (Korpora) in ausreichendem Maße frei zugänglich, können automatische Lernverfahren auf diesen Daten operieren und Sprachmodelle automatisch induzieren.

Dieser Ansatz wurde in der Computerlinguistik erfolgreich für die Alltagssprache durch die Nutzung von Nachrichtentexten aus Zeitungen und Nachrichtendiensten, digital verfügbaren Enzyklopädien (wie Wikipedia) einerseits und Texten aus sozialen Netzwerken und Medien (Twitter, Facebook, Instagram) andererseits beschritten. Für Fachtexte, mehr noch für klinische Texte stößt dieses Vorgehen jedoch an Grenzen, die durch in der EU besonders rigide Datenschutzgesetze (etwa die DSGVO)³ definiert sind.

Um die Zugänglichkeit klinischer Sprachdaten im Einklang mit den geltenden rechtlichen Rahmenbedingungen überhaupt erst zu ermöglichen, müssen potenziell individuenidentifizierende Attribute in Texten de-identifiziert werden. Mit der Annotation dieser Attribute in klinischen Texten wird hierfür die Voraussetzung geschaffen.

Die hier beschriebenen und annotierten Kategorien von PHI-Daten in Texten orientieren sich an den im US-Recht verankerten Kriterien des *Health Insurance Portability and Accountability Act* (HIPAA), wurden jedoch den deutschen Anforderungen und klinischen Besonderheiten angepasst. Durch die Bereitstellung dieser Leitlinien soll es der Forschungsgemeinschaft ermöglicht werden, einerseits unsere Daten zu reproduzieren, andererseits unsere Vorschläge in zukünftigen Projekten zu nutzen bzw. weiterzuentwickeln.

Die vorliegende Annotationsleitlinie ist Grundlage der folgenden Veröffentlichung:

(1) Kolditz et al. *Annotating German Clinical Documents for De-Identification*, MedInfo 2019 (<https://doi.org/10.3233/shti190212>).

3 Daten

Die Annotationen, für die die hier dokumentierten Leitlinien entwickelt wurden, wurden an realen klinischen Arztbriefen durchgeführt. Diese Briefe sind Teil des 3000PA-Textkorpus (2). Das Korpus wurde im Rahmen der Pilotphase des SMITH-Projekts an den Standorten Jena, Leipzig und Aachen nach folgenden Kriterien zusammengestellt:

- Alle Briefe entstanden zwischen 2010 bis 2015.
- Alle in den Briefen beschriebenen Patienten waren mindestens fünf Tage in einem der Universitätsklinika Jena, Leipzig oder Aachen stationär in Behandlung.
- Die Behandlung fand auf einer intensivmedizinischen oder internistischen Station statt.
- Alle Patienten waren zum Zeitpunkt der Datenerhebung im September 2016 bereits verstorben.

Das Jenaer Segment von 3000PA besteht insgesamt aus 1106 deutschsprachigen Arztbriefen, die sich aufteilen in:

- 383 Kurzberichte,
- 103 Verlegungsbriefe und
- 620 Entlassbriefe.

³<https://eur-lex.europa.eu/eli/reg/2016/679/oj>

4 Annotationsleitlinien

4.1 Allgemeine Hinweise

Dieser Annotationsleitlinie werden einige Begriffsklärungen vorangestellt, die zentrale Konzepte von Annotationsprozessen klären sollen:

- **Token:** kleinster Bestandteil eines Textes, der für die Annotation relevant ist, z.B. einzelne Wörter und Satzzeichen voneinander getrennt.
- **Annotat:** inhaltlich markiertes Textstück (ein einzelnes Token oder eine Spanne von Token) in einem Text, hier in einem Arztbrief.
- **Entität:** semantischer Typ eines Annotats; dieser Typ ist Teil grundlegender Taxonomien, die den Gegenstandsbereich der Annotationsaufgabe inhaltlich (kategorisch) strukturieren. Das jeweilige Annotat kann als konkrete Ausprägung (Instanz) eines Typs betrachtet werden.
 - Beispiel: Das Annotat *[01.01.2018]* wird der Entität *Date* zugeordnet.

4.2 Typensystem für PHI-Annotationen

Von den folgenden *Protected Health Information*-Typen soll immer der am spezifischsten zutreffende Typ gewählt werden, d.h., spezifischere Untertypen sind generischeren Typen vorzuziehen. (Zusammenfassung siehe Appendix A.)

4.2.1 Person

Unter diese Kategorie werden alle zu annotierenden Personennamen subsumiert; sie sind wie folgt taxonomisch differenziert:

- **Staff:** medizinisches Personal (Ärzte, Schwestern, Pfleger, ...) ⁴
- **Patient:** der im Arztbrief beschriebene behandelte Patient,
- **Relative:** Angehörige des beschriebenen behandelten Patienten.

Der allgemeine Typ **Person** sollte nur annotiert werden, wenn unklar ist, ob es sich um Personal, den Patienten oder Angehörige handelt, bzw. wenn die Person sich keiner der drei Untergruppen zuordnen lässt.

4.2.2 Date

Unter die Kategorie **Date** fallen alle Datumsangaben. Angaben von Geburtsdaten sollen gesondert mit dem Subtyp **Birthdate** annotiert werden. Der Typ **Date** soll nur verwendet werden, wenn es sich nicht um ein Geburtsdatum handelt. Doppelannotationen sollen vermieden werden.

4.2.3 Age

Mit der Kategorie **Age** werden alle Altersangaben von Patienten sowie deren Angehörigen annotiert.

⁴Wir verwenden dem sprachlichen Usus der Mehrheit der Sprechergemeinschaft des Deutschen und linguistischen Argumenten folgend das generische Maskulinum zur Referenzierung aller Geschlechter.

4.2.4 Location

Unter der Kategorie **Location** werden alle Ortsgaben und Adressen erfasst: Straße, Hausnummer, PLZ, Stadt, Kreis, Bundesland, Land und auch die Zimmernummer einer Station.

4.2.5 MedicalUnit

Mit der Kategorie **MedicalUnit** sind alle medizinischen Einrichtungen (Krankenhaus, spezielle Sprechstunden, Abteilungen, Stationen etc.) eingeschlossen. Hierzu zählen Namen spezifischer Einrichtungen, also *unsere [Klinik für Kardiologie]* und *[Gefäßsprechstunde]*, *Leiter der [Pneumologie & Allergologie]*, aber *nicht unsere Klinik/unser Haus/unsere Sprechstunde, internistisches Konsil*. Es wird hier Abteilung von Fachgebiet unterschieden: *FÄ für Innere Medizin/Allergologie* (keine Annotation für das *Fachgebiet*) vs. *Leiter der [Inneren Medizin/Allergologie]* (hier wird die *Abteilung* annotiert).

Universitätsnamen stehen in Dokumenten des 3000PA-Textkorpus i.d.R. für die entsprechenden Universitätskliniken und sind als **MedicalUnit** zu annotieren.

4.2.6 ID

Jegliche ID-Nummern oder Codes (Patienten- oder Fallnummern, IDs aus medizinischen Subsystemen, Versicherungsnummern, etc. werden durch die Kategorie **ID** markiert.

Der Subtyp **Typist** wird gewählt, um Kürzel von Schreibkräften und Sekretariaten zu kennzeichnen.

4.2.7 Contact

Mit der Kategorie **Contact** werden URLs, IP-Adressen, E-Mail-Adressen, Fax- oder Telefonnummern markiert.

4.3 Other

Mit der Kategorie **Other** werden alle schützenswerten Passagen annotiert, für die keine der obigen Kategorien zutrifft.

Namen von IT-Systemen bestimmter Hersteller, die Rückschlüsse auf die Klinik zulassen, werden ebenfalls als **Other** annotiert (nur wenn Hersteller- oder Produktnamen verwendet werden, nicht aber bei generischen Bezeichnungen), z.B. *COBRA*, *ORBIS*, *SAP*.

5 Das Annotationswerkzeug BRAT

Die folgenden Abschnitte beschreiben die Installation bzw. Konfiguration sowie die Verwendung des BRAT-Annotationswerkzeugs für die manuelle Annotation von PHI-Kategorien in deutschsprachigen Arztbriefen.

5.1 Hinweise zur Konfiguration von BRAT

Für die manuelle Annotation der Texte wird das Werkzeug BRAT RAPID ANNOTATION TOOL – BRAT (3) verwendet.⁵ BRAT wird über einen Internet-Browser verwendet und kann entweder lokal oder auf einem Server installiert werden.

Vorab muss BRAT konfiguriert werden. Die notwendigen Konfigurationsdateien werden unter <https://doi.org/10.5281/zenodo.7707882> bereitgestellt.

5.2 Hinweise zur Nutzung von BRAT

In den folgenden Abschnitten wird beschrieben, wie mit einer für die Aufgabe angepassten BRAT-Konfiguration PHI-Kategorien in deutschsprachigen Entlassbriefen manuell annotiert werden.

5.2.1 Hinweise zum Annotationswerkzeug BRAT

Im Browser muss sich jeder Annotator vorab oben rechts mit den zugewiesenen individuellen Login-Daten einloggen. Im Anschluss daran wählt man aus der *Collection* eine Aufgabe bzw. einen Text aus und kann daraufhin mit der Bearbeitung beginnen.

1. Durch einen Doppelklick auf ein Token wird dieses markiert, und ein Menü zur Auswahl aus der Menge von zulässigen PHI-Typen erscheint (s. Abb. 1). Die Markierung kann durch Klicken oder Ziehen durchgeführt werden. Zur Auswahl stehen daraufhin die Annotationstypen, mit denen ein Token oder eine Spanne von Token einer PHI-Kategorie zugeordnet werden kann.

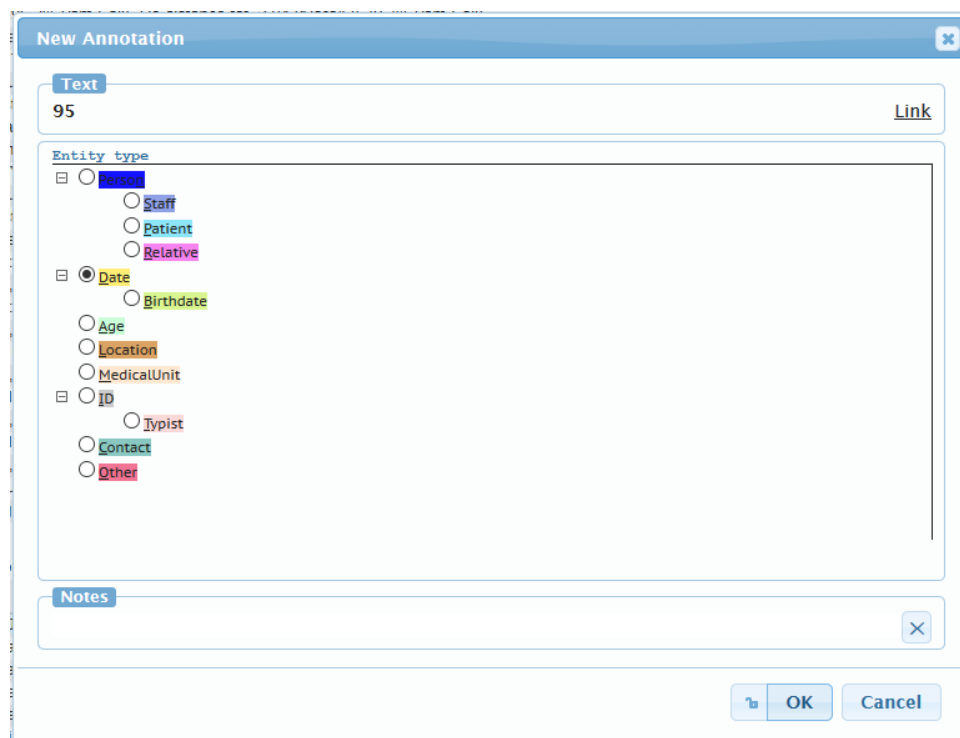


Abbildung 1: Allgemeine Auswahl nach Markierung

⁵<http://brat.nlplab.org/index.html>

2. Um die Produktivität während der Annotation zu erhöhen, hat jede Annotationskategorie zur komfortableren Auswahl ein Tastatur-Kürzel bzw. Short-Cut-Symbol (entsprechender Buchstabe in Abb. 1 unterstrichen, siehe dazu auch Appendix A).
3. Sollte ein fehlerhaft eingefügtes Annotat zu löschen sein, kann das Annotat angeklickt werden. Daraufhin öffnet sich das Auswahlmenü (siehe Abb. 2). Es werden dann mehrere Optionen angeboten:
 - *Delete*: Das Annotat wird gelöscht und kann danach einfach neu gesetzt werden.
 - *Move*: Die Spanne des Annotats kann per Maus einfach neu gezogen werden.
4. Die Funktion *Add Frag.* sollte bei der Annotation der PHI-Kategorien nicht zum Einsatz kommen.

Abbildung 2: Korrektur eines Annotats

5.2.2 Inhaltliche Hinweise zur Annotation

Eine Annotation sollte jeweils genau eine Entität abdecken (**1 Annotation : 1 Entität**).

Ausnahmen: Alle Namen und Kürzel der Schreibkräfte bzw. Ersteller der Niederschrift werden zusammen annotiert, z.B. [müll/mei/gau].

- Wenn mehrere Entitäten unmittelbar aufeinanderfolgen, müssen diese getrennt annotiert werden, z.B. *Dr.med. [Max Mustermann], Prof. Dr. [Maria Mustermann], Dres. [Müller]/[Meier]* → jeweils zwei getrennte *Staff*-Annotationen.
- Medizinische Untereinheiten und Orte bekommen einzelne Annotationen, auch wenn sie direkt aufeinanderfolgen, z.B. *[XY-Station] der [XY-Klinik], [XY-Stadt]*,
 - **Ausnahme:** wenn es sich um einen feststehenden Namen handelt, der aus mehreren Bestandteilen besteht, z.B. *[St. Lioba-Klinikum zu Fulda]*.

Folgende Konventionen sind zudem zu beachten:

- **Adressbestandteile** werden *einzel*n nach folgendem Muster annotiert:
[*Straße Hausnummer*], [*PLZ Ort*], wobei die Reihenfolge natürlich abweichen kann.
- Personen werden immer **ohne Anrede und Titel** annotiert, z.B.
Frau Prof. Dr. [Maria Mustermann], auch abgekürzte Namen und Initialen werden annotiert.
- Altersangaben **exkl. Jahre** (nur die Zahl), z.B. [*80*] *Jahre*, [*80*]-*jährige*.
- **Zeitspannen** werden in einzelne Datums-/Zeitangaben **aufgebrochen**, z.B. [*01.01.2018*]
bis [*31.12.2018*] → zwei *Date*-Annotationen.
- Stationsnummern werden immer **inkl. Station** annotiert, z.B. [*Station 105*].
- *Date* ohne zusätzliche qualifizierende Angaben wie *Anfang/Mitte/Ende* [*Oktober*].

Anmerkung: Die hier folgenden Instruktionen setzen (automatisierte) Vorannotationen voraus, z.B. von Datumsangaben. Dementsprechend bestehen zwei parallele Aufgaben:

1. Fehlende Annotationen sollen ergänzt und überflüssige gelöscht werden.
2. Die Spannen der Annotationen sollen korrigiert und vereinheitlicht werden, sodass alle sensiblen Informationen abgedeckt werden und eine möglichst hohe Übereinstimmung zwischen den Annotationen (Inter-Annotator-Agreement, IAA) erreicht wird.

Workflow: Es sollen zunächst die Vorannotationen ignoriert werden – prioritär soll nach Elementen gesucht werden, die nicht annotiert wurden, und ggfs. ergänzt werden müssen. Danach sollen alle Annotationen durchgegangen und entschieden werden, ob sie notwendig sind. Ggfs. muss jeweils die Spanne und der automatisch vorannotierte Typ modifiziert werden.

5.2.3 Was nicht zu annotieren ist

Um nicht zu viele Informationen zu verlieren und um eine möglichst hohe Übereinstimmung zwischen den einzelnen Annotationen zu erreichen, ist es wichtig, konsistent zu annotieren. Folgende sprachliche Ausdrücke sollen **nicht Teil von Annotationen** sein:

- **Funktionswörter** wie Präpositionen (z.B. *von, mit, aus, zu, in*), Artikel (z.B. *der, die, das, ein*), Pronomen (z.B. *unser, sein, sie*), Konjunktionen etc. **am Rand** (am Anfang oder Ende einer Annotation)
 - **Ausnahme:** Namensbestandteile sind Teil der Annotationsspanne (z.B. [*von der Leyen*], nicht aber das Vorfeld einer Präpositionalphrase (hier nur den nominalen Kopf annotieren): *auf der* [*HNO-Station*])
- **Kategorisierungen, Attribute und nähere Beschreibungen** eines Individuums (z.B. *geboren am, wohnhaft in*)
- **Appellativa**, die keine Rückschlüsse auf sensible Informationen zulassen (z.B. **Kollegen** *der* [*HNO*])
 - **Ausnahme:** Komposita wie [*Intensivstation*], [*Laborabteilung*], [*HNO-Station*] (hier jeweils das gesamte Token annotieren) und Ausdrücke, die medizinische Einrichtungen referenzieren, z.B. [*Klinik für Kardiologie*], [*Klinik für AVG-Chirurgie*]

- **Verwandtschafts- und Affinitätsbezeichnungen** ohne Namen (z.B. *Tochter, Mutter, Großvater, Ehefrau*)
- **Appellativa** (Gattungsbegriffe), die allein nicht auf eine bestimmte Entität verweisen (z.B. *Kollege, Patient, Oberarzt, Notarzt, Allgemeinmediziner, Praxis*)
 - **Ausnahme:** Begriffe, die für bestimmte medizinische Einrichtungen, Stationen, Sprechstunden stehen, z.B. *[Klinik für Kardiologie], [Gefäßsprechstunde]*
- **Positionen oder Berufsbezeichnungen** von **Ärzten/Personal**, z.B. *Oberärztin Frau Dr. [Musterfrau], Direktor der Klinik, Leiter der [HNO]* (durch die Anonymisierung der Abteilung ist *Leiter* nicht mehr identifizierend)
- **Pronomen**, z.B. *er, sie*
- **Tageszeitangaben**, z.B. *12:32, 6:30 Uhr*, einzelne **Wochentage ohne Monat oder Jahr**, z.B. *jeden Freitag, letzten Montag, den ganzen Sonntag, mo/di* (z.B. bei Medikationsangaben)
- **Ima-** und **Seriennummern**, z.B. *Serie 600, Ima 30*
- **Konsile, Boards** (hier handelt es sich nicht um feste medizinische Einheiten, die einen wiedererkennbaren Namen haben könnten)

A Appendix

Typ	Beschreibung	Short-Cut
Person	Personennamen	N
Staff	medizinisches Personal (Ärzte, Schwestern, Pfleger)	S
Patient	behandelter Patient	P
Relative	Angehörige des Patienten	R
Date	jegliche Datumsangabe	D
Birthdate	Geburtsdatum	B
Age	Altersangaben (Patient oder Angehörige)	A
Location	Ortsangaben (Adressen: Straße, Hausnummer, PLZ, Stadt, Kreis, Bundesland, Land; Zimmernummer)	L
MedicalUnit	medizinische Einrichtungen (Krankenhaus, spez. Sprechstunden, Abteilungen, Stationen etc.)	U
ID	ID-Nummern oder Codes (Patienten- oder Fallnummern, IDs aus medizinischen Subsystemen, Versicherungsnummern)	I
Typist	Kürzel der Schreibkräfte	T
Contact	URLs, IP-Adressen, Email-Adressen, Fax- oder Telefonnummern	C
Other	schützenswerte Information, die von keiner der anderen Kategorien subsumiert wird	0

Tabelle 1: Übersicht der Typensystems für PHI-Annotationen und deren Short-Cuts der BRAT-Konfiguration

Referenzen

- [1] Tobias Kolditz, Christina Lohr, Johannes Hellrich, Luise Modersohn, Boris Betz, Michael Kiehntopf, and Udo Hahn. Annotating German clinical documents for de-identification. In Lucila Ohno-Machado and Brigitte Séroussi, editors, *MEDINFO 2019 — Proceedings of the 17th World Congress on Medical and Health Informatics: Health and Wellbeing e-Networks for All. Lyon, France, 25-30 August 2019*, number 264 in Studies in Health Technology and Informatics, pages 203–207, Amsterdam, Berlin, Tokyo, Washington, D.C., 2019. IOS Press.
- [2] Udo Hahn, Franz Matthies, Christina Lohr, and Markus Löffler. 3000PA: towards a national reference corpus of German clinical language. In Adrien Ugon, Daniel Karlsson, Gunnar O. Klein, and Anne Moen, editors, *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth - Proceedings of MIE 2018, Medical Informatics Europe, Gothenburg, Sweden, April 24-26, 2018*, volume 247 of *Studies in Health Technology and Informatics*, pages 26–30. IOS Press, 2018.
- [3] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. BRAT: A Web-based tool for NLP-assisted text annotation. In Frédérique Segond, editor, *EACL 2012 — Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations. Avignon, France, April 25-26, 2012*, pages 102–107, Stroudsburg/PA, 2012. Association for Computational Linguistics (ACL).