Annotationsrichtlinien - <add corpus name>

Leila Feddoul, Clara Lachenmaier, Sarah Bachinger

30. März 2023

1 Einleitung

1.1 Kontext

Um eine Verwaltungsleistung (z. B. die Autoanmeldung) für Bürger und Unternehmen erbringen zu können, müssen die Behörden einen bestimmten Prozess durchführen. Dieser Prozess wird nicht willkürlich erstellt, sondern beruht in der Regel auf Rechtsgrundlagen (z. B. Gesetze, Verordnungen usw.). Der erste Schritt bei der Erstellung von Verwaltungsprozessen besteht also darin, relevante Rechtsgrundlagen zu sammeln und diese in einem weiteren Schritt zu analysieren. Ziel dieser Analyse (Normenanalyse) ist es, Hinweise auf mögliche Prozesselemente (z.B. Akteure, etc.) im Text zu identifizieren und zu klassifizieren (annotieren) durch die Zuordnung einer bestimmten Kategorie zu einem entsprechenden Wort/einer Folge von Wörtern. Diese Annotation wird manuell vorgenommen, was sehr zeitaufwändig ist. Aus diesem Grund wollen wir ein System entwickeln, das relevante Informationen automatisch erkennt und annotiert. Eine der Techniken, die zur Lösung dieser Aufgabe verwendet wird, ist das überwachte Lernen, für das annotierte Trainingsdaten (Trainingskorpus) benötigt werden. Menschen müssen also eine gewisse Menge an Daten manuell annotieren, die verwendet werden, um Algorithmen des maschinellen Lernens beizubringen, dieselbe Aufgabe auf neuen Texten automatisch durchzuführen.

1.2 Daten

Die derzeit verfügbaren Daten sind eine Kollektion von Gesetzen, die von der Website gesetze-im-internet.de über das fimportal.de gesammelt wurden. Das letztere ermöglicht eine Suche über vorhandene Verwaltungsleistungen, und bietet eine strukturierte Beschreibung jeder Leistung (z. B. Kurztext) inklusive einem Feld namens Rechtsgrundlagen", bestehend aus einer Liste von Links zu den Gesetzen, die als Grundlage für die jeweilige Leistung dienen. Dadurch war es möglich, einen unannotierten Korpus zu erstellen, der aus einer Anzahl von Textdokumenten besteht, wobei jedes Textdokument dem Gesetzestext einer bestimmten Dienstleistung entspricht.

2 Aufgabenbeschreibung

Dieser Abschnitt gibt einen Überblick über die verschiedenen Kategorien, die für die Annotation der gesammelten Texte verwendet werden sollen. Das generelle Ziel der Normenanalyse ist es, Hinweise im Text zu finden, die es ermöglichen, die notwendigen Schritte zu identifizieren, die für die Erbringung einer Leistung erforderlich sind. Zu diesem Zweck definieren wir folgende Kategorien:

Hauptakteur. Aus Verwaltungssicht die Person oder Behörde, die für die Leistungserbringung verantwortlich ist (*Wer ?*). *Beispiel*: Bundesamt für Sicherheit in der Informationstechnik.

Ergebnisempfänger. Die Person oder das Unternehmen, das die Verwaltungsleistung in Anspruch nehmen möchte (*Wen ?*) . *Beispiel*: Antragsteller.

Mitwirkender. Externe Behörde oder eine andere Rolle (z.B. Betreuer) die bei der Verwaltungsleistung mit einbezogen wird (*Mit wem ?*). *Beispiel*: Bundesministerium des Innern, für Bau und Heimat.

Aktion. Handlung der Verwaltung oder der anderen Akteure, z.B. Antragssteller, extern Mitwirkender etc. (*Was ist prozessrelevant? Was wird getan?*). *Beispiel*: beantragen, vorlegen, erteilen etc.

Signalwort. Angabe zur Verbindlichkeit einer Aktion. Ausgedrückt durch Modalverben oder andere Begriffe die beschreiben, wie optional eine Aktion ist. *Beispiel*: bei Bedarf, erforderlich, zweckgebunden, auf Verlangen, muss soll, kann etc.

Dokument. Dokumente die zwischen Akteur*innen ausgetauscht werden (z.B., von Behörde verlangt) (Gegenstand der Aktion:*Womit ?*). *Beispiel*: Sicherheitszertifikat, Antrag.

Bedingung. Voraussetzung für die Aktion, z.B. zeitlicher Auslöser (*Welche Voraussetzung?*). Gewisse Wörter "wie wenn, soweit, sofernäber auch jegliche Verben wie in „sind ...“oder „verstirbt .. “, können Hinweis auf Bedingungen sein. *Beispiel*: Die Anträge werden in der zeitlichen Reihenfolge ihres Eingangs bearbeitet; hiervon kann abgewichen werden, *[wenn das Bundesamt wegen der Zahl und des Umfangs anhängiger Prüfungsverfahren eine Prüfung in angemessener Zeit nicht durchführen kann und an der Erteilung eines Zertifikats ein öffentliches Interesse besteht]*.

Frist. Fristen oder Zeitlimits, die für den Prozess relevant sind oder die sich auf einen Dokument beziehen, aber auch andere Fristen. *Beispiel*: Die Bundesagentur entscheidet über vollständige Anträge *[innerhalb von sechs Wochen]*.

Datenfeld. Bei expliziter Nennung von Elementen, die mit in einen Antrag müssen. *Beispiel*: In der Anzeige sind anzugeben: 1. der *[Zweck des Eingriffs]*, ... und voraussichtliche *[Dauer des Vorhabens]*.

Handlungsgrundlage. Verweise auf andere Gesetze. *Beispiel*: Absatz 4, Absatz 1 Satz 2 Nr.3, Satz1 Nummer 8 Buchstabe d.

Abbildung 1 zeigt einige Beispiele von annotierten Dokumententeile unter Verwendung der zuvor definierten Kategorien im Annotationstool INCEpTION.

Details zur Toolverwendung und zu den Annotationsrichtlinien (Annotationsprozess sowie grundlegende und kategoriespezifische Regeln) werden in den folgenden Abschnitten erläutert.

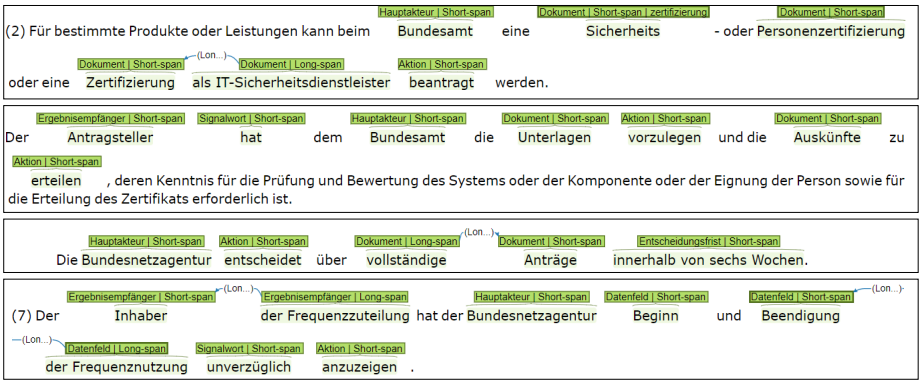

Abbildung 1: Beispiele von annotierten Dokumententeile unter Verwendung der zuvor definierten Kategorien im Annotationstool INCEpTION.

Ein entscheidender Aspekt bei der Erstellung von Trainingskorpora mit menschlichen Annotationen sind präzise und umfassende Annotationsrichtlinien. Sie definieren die Aufgabe genauer und haben zum Ziel, eine konsistente Annotation durch verschiedene Annotatoren zu gewährleisten. Dadurch wird auch verhindert, dass die Modelle verwirrt werden und eine geringe Trefferquote erzielen, weil sie mit inkonsistenten/veränderlichen Annotationen versorgt werden, aus denen sie nur schwer lernen können.

3 Annotationstool Inception

Für die Annotation wird eine Instanz der Plattform Inception, Version 26.0, genutzt[**inception2018**]. Diese ist online erreichtbar [1], wobei der Zugriff Login-Daten erfordert. Im Folgenden wird eine kleine Einführung basierend auf dem User Guide[2] von Inception in die Annotation mit Inception gegeben.
Nach dem Login wird auf die Projektübersichtsseite weitergeleitet. Dort ist es möglich, mehreren Projekten zugeordnet zu sein. Im Normalfall wird die Rolle „Annotator“ vergeben.
Ein Annotationsprojekt besteht aus mehreren Dokumenten, die gemäß dieser Annotationsguidelines annotiert werden sollen.
Die Ansicht der Annotationsseite ist in Abbildung 2 zu sehen. Innerhalb eines zu annotierenden Dokuments können einzelne Wörter durch Doppelklick oder Markieren eines oder mehrerer Wörter annotiert werden.
Auf der rechten Seite können dann ein FIM-Tag, Span-Typ, Prefix und Suffix angegeben werden. Verpflichtend sind FIM-Tag und Span-Typ. Die Auswahlmöglichkeiten werden in späteren Sektionen definiert.
Zwischen zwei Annotationen können durch eine Relation mithilfe von Ziehen

[1] `https://canareno-annotation.fmi.uni-jena.de`
[2] `https://inception-project.github.io/releases/26.0/docs/user-guide.html`

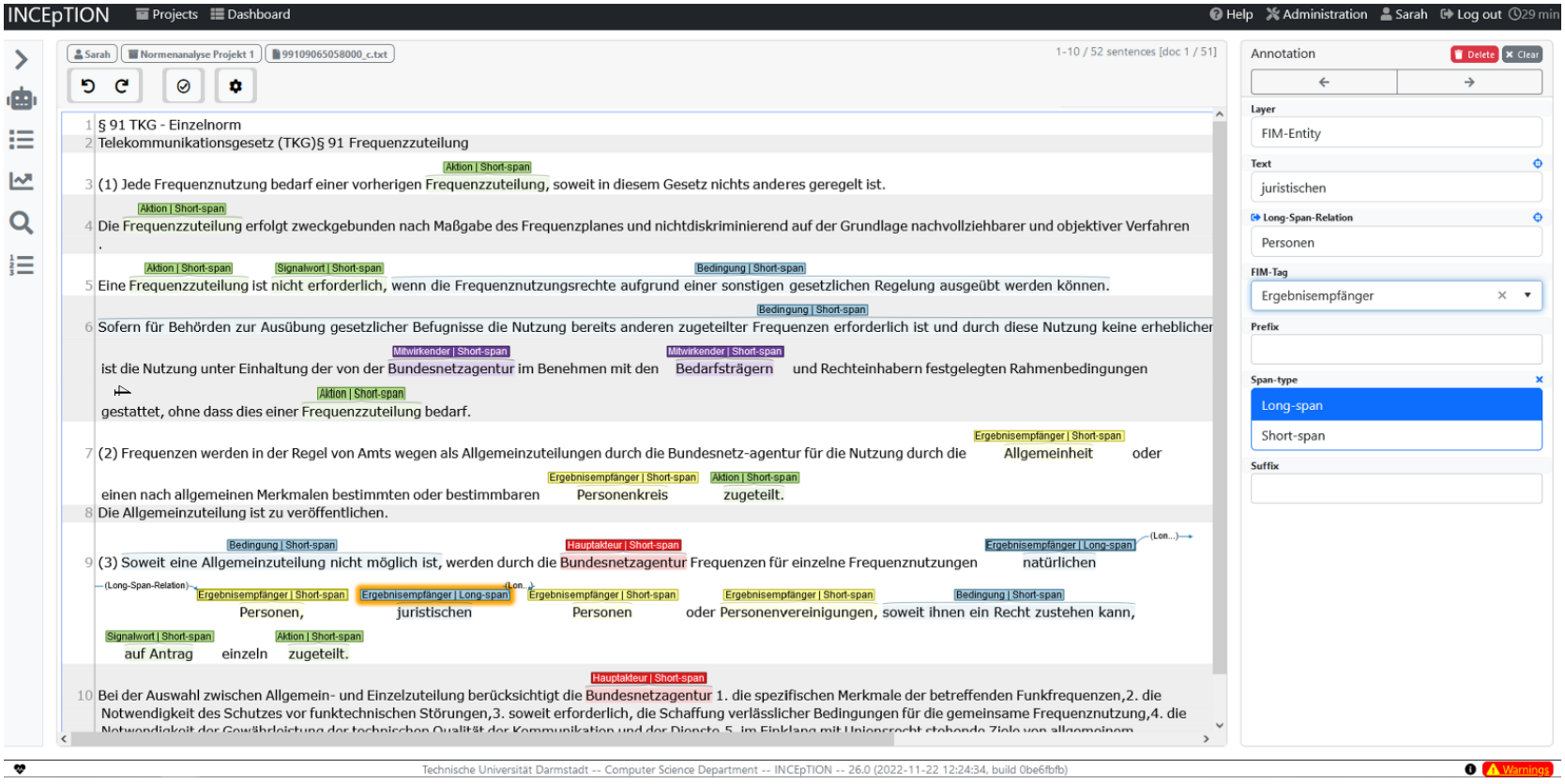

Abbildung 2: Ansicht der Inception-Seite für die Annotation eines Textes.

des Mauszeigers von der einen Annotation zur anderen verbunden werden. Diese wird dann auf der rechten Seite als zusätzliches Feld angegeben.
Graue hinterlegte Annotationen sind automatisch generierte Vorschläge des Stringmatchers, der aus vorherigen Annotationen neue lernt. Die Vorschläge können entweder durch einen einfachen Klick angenommen oder durch einen Doppelklick abgelehnt werden.
Oberhalb des Dokuments befindet sich eine Reihe von Kacheln. Die ersten Zwei realisieren „Aktion rückgängig machen“ und „Rückgängig machen zurücknehmen“. Mit Klick auf den Kreis mit dem Haken wird ein Dokument als beendet markiert und es wird ein neues Dokument zugewiesen. Es ist zu beachten, dass es nur Personen mit den Rollen „Manager“ oder „Curator“ ein beendetes Dokument wieder reaktivieren können. Die letzte Kachel öffnet Einstellungen zum aktuellen Annotationseditor, wie die Farbenvariation der Annotationen.

4 Annotation Guidelines

4.1 Annotationsprozess

Vor der Annotation eines neuen Dokuments sollte jeder AnnotatorIn die zugehörige Leistungsbeschreibung auf *https://fimportal.de/detail/L/\{document_titel\}* öffnen und gründlich durchlesen, um sich den Kontext zu erarbeiten.

Hinweis: Auf die Leistungsbeschreibung lässt sich nur im eingeloggten Zustand zugreifen.

Als erster Schritt mit einem zu annotierenden Dokument, soll dieses zunächst einmal komplett durchgelesen werden. Danach sollen alle Bestandteile, die sich einem Tag aus dem Tagset (siehe Sektion 2) zuordnen lassen gefunden und annotiert werden. Dabei sollte jedoch genau darauf geachtet werden, ob das gefundene Kandidat auch für die Leistung, die bearbeitet wird, relevant ist.

Die Annotationsplattform INCEpTION stellt ein Tool zur Verfügung, den **String Matcher**, der Wörter, die bereits an einer anderen Stelle im Text getaggt wurden, automatisiert hervorhebt (grau) und das zuvor gewählte Annotationstag vorschlägt. Durch Klicken auf den Vorschlag, kann dieser angenommen werden. Bei der Annahme von String-Matcher-Vorschlägen, ist ebenfalls stets darauf zu achten, dass diese relevant für die Leistung sind.

Die folgenden Informationen sollten notiert werden: die geschätzte Zeit, die für die Annotationen der einzelnen Dokumente benötigt wird, sowie alle Schwierigkeiten und offenen Fragen, die später besprochen werden sollen. Schon während des Annotationsprozesses sollten die Daten regelmäßig exportiert werden, um Fortschrittsverlusten durch z.B. abstürzen des Browsers vorzubeugen.

4.2 Grundlagen

4.3 Annotationsbereich

Nur Abschnitte, die für die Leistungsbeschreibung relevant sind, sollten auch annotiert werden. Ein Hinweis dafür bietet auch die Angabe des relevanten Absatzes/Paragraphs, die sich manchmal in der Leistungsbeschreibung (Rechtsgrundlage, Beispiel) findet. Die Angabe des Absatzes schließt aber oft nichts aus, dass auch in anderen Absätzen relevante Kandidaten stehen und ersetzt deshalb nicht das aufmerksame Lesen und Annotieren der restlichen Absätze.

4.4 Allgemeine Regeln

- Titeln werden nicht annotiert Beispiel: Gesetz über das Bundesamt für Sicherheit in der Informationstechnik
- Nomen werden ohne Artikel (*der, die, das ..*) annotiert.
- Die Pluralformen von Wörtern (*Anträge* vs. *Antrag*) werden annotiert.
- Ein Wort, das mit einem Bindestrich verbunden wird (*schwarz-weiß*) und Fremdwörter und Abkürzungen (*BMI*) werden als ein Wort annotiert.

- Falsch geschrieben Wörter (*Haupt akteur*) werden nicht annotiert.
- Würden sich zwei Annotationen überlappen, wird die flächenmäßig größere Annotation gewählt.
- Wenn Wörter in anderen Sprachen auftauchen, sollten auch diese entsprechend annotiert werden.

4.5 Spezifische Regeln

4.5.1 Long-/short span Annotationen

INCEpTION erlaubt eine Relationsannotation: eine Annotation, die die Zusammengehörigkeit einzelner Teile anzeigt. Dieses Feature erlaubt es, mittels einer einzigen Annotation zwei Korpora zu generieren (Abbildung 3). Nehmen wir das *Dokument Bescheinigung über die Wohnberechtigung*. Im Normalfall müsste die Entscheidung getroffen werden, ob wir ein einfaches Korpus mit nur dem Kernstück der obigen Phrase *Bescheinigung* oder der ganzen Phrase *Bescheinigung über die Wohnberechtigung* erhalten wollen.

Korpus	Annotiert
Simple	Bescheinigung
Enriched	Bescheinigung über die Wohnberechtigung

Abbildung 3: Überblick über geplante Datensätze mit einer Beispiel Annotation

Wir umgehen diese Entscheidung, indem bei jedem Tag angegeben werden muss, ob es sich um eine Shortspan (Bestandteil des Simple und des Enriched Corpus) oder eine Longspan (Bestandteil des Enriched Corpus) handelt. Longspaninstanzen können nicht alleine stehen und müssen mit einer Shortspaninstanz mit dem gleichen Label via Spanrelation verbunden werden (Abbildung 4). Jeder Entität sollte ein Span-type (Short, Long) zugewiesen werden (Pflichtfeld).

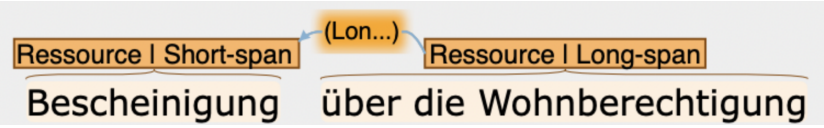

Abbildung 4: Beispiel: Long-Short-Span Annotation

Generell sollte so minimal wie möglich annotiert werden, z.B *Die Bescheinigung über die Wohnberechtigung, die in bis zu zwei Wochen einzureichen ist.* Auch wenn "die in bis zu zwei Wochen einzureichen ist"die Bescheinigung näher beschreibt sollte nur "über die Wohnberechtigungäls Zustatz annotiert werden. Im folgenden werden Anwendungsfälle für die Long-/Shortspanannotation gezeigt:

- **Nomen + Adjektiv**: *Medizinisch-technischer* **Long-span** *Laboratoriumsassistent* **Short-span**
- **Nomen + Nomen**: *Agentur* **Short-span** *für Arbeit* **Long-span**
- **Nomen + Nomen in Genitivkonstruktionen**: *Vereinbarung* **Short-span** *eines Termins* **Long-span**
- **Nomen + Verb**: *Führerschein* **Long-span** *beantragen* **Short-span**

Weitere Sonderregeln.

- *Wie behandelt man Kategorien, die andere Kategorien enthalten ?*: wenn eine lange Spanne andere Kategorien enthalten würde, nicht die lange Spanne annotieren, sondern zwei kurze Spannen erstellen. *Beispiel:* Prüfung [Aktion, Short] der Sachkunde [Dokument, Short].
- *Wie sind generische Hinzufügungen von langen Spannen zu behandeln ?*: generische Hinzufügungen wie *unterschiedliche* in *unterschiedliche Dokument* werde nicht annotiert, weil sie sehr allgemein sind.
- *Wie werden Verben mit zweiteiliges Zeitformen annotiert (z.B. beantragt werden, hat beantragt)?*: Nur das Vollverb soll annotiert werden.

4.5.2 Ellipsen

Um dem Informationsverlust bei elliptischen Konstruktionen vorzubeugen, gibt es Präfix- und Suffixfelder. Ein elliptischer Zusammenschluss sind Phänomene wie *Sicherheits- und Personenzertifizierung.* Es handelt sich hierbei um zwei inhaltlich voneinander unabhängige Entitäten, die aber wegen der Auslassung von *zertifizierung* im Falle von Sicherheit. Diese Auslassung wird im Sprachgebrauch mitgedacht. SprecherInnen wissen intuitiv, dass es sich um eine Sicherheitszertifizierung handelt. In der Schriftsprache wird dieses Wissen noch durch ein - verstärkt. Bei einem Computer kann dieses Mitwissen nicht vorausgesetzt werden.

In einem solchen Fall sollten die beiden Teile als zwei Instanzen behandelt werden unter Einhaltung folgender Regeln:

- Der Übersichtlichkeit halber werden **nur** Fälle, in denen die Ellipse durch einen - und ein verbindendes Element wie *und* oder *oder* zwischen sich haben, berücksichtigt.

- Der verbindende Terminus (*oder*, *und*, und Bindestrich etc.) sollte nicht annotiert werden.
- Die optionale Felder **Suffix** oder **Präfix** sollen manuell ergänzt werden.
- Befindet sich die Auslassung (und der Bindestrich) am Ende eines Wortes muss das Attribut *Suffix* manuell durch den ausgelassenen Begriff (zertifizierung, siehe Abbildung 5) eingetragen werden.
- Befinden sich Auslassung und Bindestrich am Anfang eines Wortes muss das Attribut *Präfix* manuell eingetragen werden (z.B Obstanbau und-ernte).
- Präfix soll immer mit Großbuchtabe anfangen und Süffix mit einem kleinen.

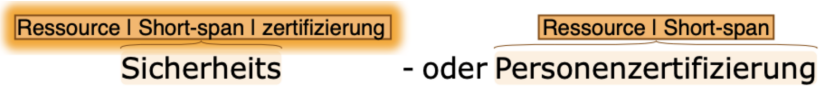

Abbildung 5: Beispiel: Ellipse

4.6 Kategorie-spezifisch

4.6.1 Signalwort

Für Signalwörter gelten noch folgende Regeln:

- Signalwörter werden annotiert wenn sie Teil der Verbform sind und signalisiert gleichzeitig die Modalität z.B. *Vertrag soll abgeschlossen werden.*
- Auch Wörter, die die Modalität beschreiben und nicht Teil der Verbform sind, wie "bei Bedarf"(offener Satz), sollten annotiert werden.
- Für Verben gibt es eine definierte geschlossene Menge, nämlich [müssen, sollen, können, dürfen]. Diese sollten nur annotiert werden, wenn sie zusammen mit einem Vollverb genannt werden Auch wenn ein Modalverb ein anderes Verb modifizieren sollte, müssen beide Verben nicht zwingend nebeneinander stehen, um annotiert werden zu können: können vor Ort bestellt werden.
- Wörter, die Signalwörter näher bestimmen wie in *braucht nicht, können nur*, sollen per Longspan dem Signalwort zugeordnet werden.
- Das Verb *sein* und *haben* wird nur in Kombination mit einem anderen Verb im zu-Infinitiv als Signalwort annotiert (z.B. hat zu entscheiden)

4.6.2 Aktion

Für Aktionen gelten noch folgende Regeln:

- Nur Aktionen aus "Verwaltungssichtöder auch Aktionen, die von z.B. Äntragsteller"durchgeführt sind, werden annotiert.
- Aktionen sind sowohl Verben *beantragt* als auch Substantive *Beantragung*.
- Im Fall von trennbare Verben wie *sicherstellen* in *Das Bundesamt stellt bis 5 Tage nach Einreichung des Antrags sicher, dass...* soll nur der flektierbare Teil des Verbs, wie in diesem Fall *stellen*, annotiert werden.
- Der sogenannte zu-Infinitiv wird manchmal innerhalb des Verbs, manchmal davor gebildet. Z.b. *einzuhalten* vs. *zu beantworten*. In zweiterem Fall wird nur das Verb und nicht die Partikel *zu* annotiert.
- Ist dem Verb ein Objekt, das mit "Gegenstand der Aktionßuzuordnen, das hilft, die Aktion näher zu beschreiben, ohne einem der anderen Annotationstags zugeordnet werden zu können, so wird dieser Gegenstand als Longspan-Aktion annotiert. Beispiel: *Frequenzen zugeteilt*

4.6.3 Bedingung

Eine Bedingung gibt Voraussetzungen für eine Aktion und werden zum Beispiel durch Wörter wie *wie wenn*, *soweit* oder *sofern* eingeleitet. Für Bedingungen gelten noch folgende Regeln:

- Als Bedingung soll jeweils der ganze konditionale Nebensatz bis zum Punkt oder Komma annotiert werden. Die Satzzeichen, die den konditionalen Nebensatz einschließen, müssen nicht annotiert werden.
- Gibt es mehrere Nebensätze als Bedingungen, sollen diese als eine einzige Bedingung annotiert werden. Beispiel: *Verstirbt eine natürliche Person, ohne dass ein Erbe die Frequenzen weiter nutzen will ...* Beispiel von Listen von Bedingungen: *wenn 1. ... , 2. ....*
- Es wird auch kein Unterschied zwischen einer Bedingung, die im Prozess entscheidend ist, und einer Voraussetzung, die vorher erfüllt sein muss, gemacht. Beides wird mit dem Label „Bedingung“ annotiert. Beispiel Bedingung: *Das Amt kann Ausnahmen zulassen, <wenn zu erfüllende Bedingung>* Beispiel Voraussetzung: *Falls ein Abschluss vorhanden ist, ...* oder *nur Personen mit abgeschlossenem Studium...*
- Außerdem zählen als einleitendes Worte wie *insoweit* für eine Bedingung, es wird aber nur der konditionale Nebensatz annotiert.
  Beispiel: *... nur insoweit erteilt werden, <konditionaler Nebensatz, der als Bedingung annotiert wird>*

- Als Bedingung zählen auch Instanzen, in denen die Bedingung auch ohne konditionalen Nebensatz aufgezeigt wird. Beispiel: *nur Personen mit abgeschlossenem Studium...*

4.6.4 Handlungsgrundlage

Das Label „Handlungsgrundlage“ soll auf alle Teile einer Rechtsgrundlage verweisen, das heißt auf den Namen des Buches zusammen mit allen zugehörigen Paragraphen und Absätzen. Dies gilt auch, wenn zum Beispiel einzelne Absätze mit „und“ getrennt sind.
Beispiel: „BGB § 4 und §5 Abs. 3, 4 und 5“: alles soll als Handlungsgrundlage annotiert werden.

4.6.5 Dokument

Für Dokumente gelten noch folgende Regeln:

- In einigen Fällen haben wir einige Objekte, die Gegenstand der Aktion sind (womit?), aber keine Dokumente sind. Diese Fälle sollten nicht als Dokumente annotiert werden, z. B. in *Derjenige, der einen Eingriff nach Absatz 1 Satz2 Nummer 4 durchführen will, hat den Eingriff spätestens zwei Wochen vor Beginn der zuständigen Behörde anzuzeigen* soll *Eingriff* nicht als Dokument annotiert werden.

- Allgemeine begriffe wie : *Auskunft* und *Unterlagen* sollen auch als Dokument annotiert werden.

4.6.6 Ergebnisempfänger

Für Ergebnisempfänger gelten noch folgende Regeln:

- Demonstrativpronomen (*derjenige*) und Relativpronomen (*wer*) werden in Gesetzen oft verwendet, um auf bestimmte Personen zu verweisen. Diese Ausdrücke werden jedoch als zu allgemein angesehen und sollten nicht als z. B. Ergebnisempfänger vermerkt werden. Zum Beispiel, in *Derjenige, der einen Eingriff nach Absatz 1 Satz2 Nummer 4 durchführen will, hat den Eingriff spätestens zwei Wochen vor Beginn der zuständigen Behörde anzuzeigen* soll *derjenige* nicht annotiert werden.

4.6.7 Frist

Für Fristen gelten noch folgende Regeln:

- Bei Fristen ist der gesamte relevante Satz zu annotieren, z.B. *bis zu vier Wochen.*