A Examples for annotating named entities in BIOfid corpus

1. *[Coeloglossum viride]TAXON blüht im [Mai]TIME auf den Wiesen besonders der [Grabenwiese]LOC vor dem [Eschenheimer Tor]LOC, wo ich sie seit [1729]TIME im [Mai]TIME fand gross und klein mit ganz grünen Blumen, auch mit einem dunkelroten Bart, mit breiten und schmalen Blättern. Einige haben auch einen Geruch, andere nicht. Zwischen [Falkenstein]LOC und [Cronberg]LOC auf Wiesen.*

2. *Das folgende stellt einen Versuch dar, aus dem Verlauf der [nacheiszeitlichen]OTH Ausbreitung der [Weissbuche]TAXON oder [Hainbuche]TAXON ([Carpinus betulus]TAXON) in [Norddeutschland]LOC einen Beitrag für die Beurteilung des Klimas namentlich zur [späten Wärmezeit]OTH zu gewinnen.*

3. *Während der Ausgrabung einer grösseren [bandkeramischen Siedlung]OTH bei [Bracht]LOC nördlich [Marburg]LOC wurde Herr Dr. [O. UENZE]PER, [Amt für Bodenaltertümer]ORG, [Marburg]LOC, auf ein der Fundstelle unmittelbar benachbartes kleines Moor von etwa 50m Durchmesser aufmerksam und vermutete, dass es die Wasserstelle der [Neolithiker]OTH gewesen sei.*

4. *[Falco]TAXON [Linnaeus]PER [1758]TIME*

5. *[Carex praecox [Jacq.]PER var. distans]TAXON [Appel]PER*

6. *Verfasser untersuchte die Winterknospen von [S. lanata]TAXON, [glauca]TAXON, [lapponum]TAXON, [phylicifolia]TAXON, ...*

7. *[Zug von [Falco vespertinus]TAXON durch [Westeuropa]LOC im [September 1927]TIME.]OTH [Ornithol. Monatsber.]ORG 36 ([1928]TIME) S.42-44.*

B Annotation Guidelines (biologized version of Benikova et al. (2014))

Guidelines für die Named Entity Recognition. Sie bauen auf den Guidelines in den STTS-Guidelines (Schiller et al., 1999), (Telljohann et al., 2012) und (Chieu and Ng, 2002) auf.

B.1 Einführung: Named Entity Recognition

Unter der Named Entity Recognition (NER) versteht man die Aufgabe, Eigennamen (named entities) in Texten zu erkennen. Technisch gesehen sind hierzu zwei Schritte notwendig. Zuerst müssen in einem laufenden Text die Token gefunden werden, die zu einem Eigennamen gehören (Named Entity Detection: NED), danach können diese Eigennamen semantischen Kategorien zugeordnet werden (Named Entity Classification). Prototypisch ist dabei der Unterschied zwischen Eigennamen und Appellativa der, dass letztere eine Gattung oder eine Klasse beschreiben, während erstere einzelne Individuen oder Sammlungen von Individuen unabhängig von gemeinsamen Eigenschaften bezeichnen (Burkhardt, 2004). Die vorliegenden Guidelines sollen es Annotatoren ermöglichen, Eigennamen in Texte aus Standard und Nichtstandard-Varietäten konsistent zu annotieren. In diesen Guidelines werden die beiden Aufgaben der NED und NEC nicht unterschieden, da die Konzentration auf Beispiele in diesem Dokument, die Trennung künstlich erzeugen müsste und nicht zu erwarten ist, dass die Resultate sich dadurch verbessern würden. In Anlehnung an die oben genannten Guidelines für Zeitungssprache werden in NoSta-D-BIOfid sechs semantische Hauptklassen unterschieden (Personen, Taxa, Organisationen, Orte, Zeiten und Andere).

B.2 Wie finde ich eine NE?

Schritt 1: Nur volle Nominalphrasen können NEs sein. Pronomen und alle anderen Phrasen können ignoriert werden.
Schritt 2: Namen sind im Prinzip Bezeichnungen für einzigartige Einheiten, die nicht über gemeinsame Eigenschaften beschrieben werden.
Beispiel:
[Der Struppi] folgt [seinem Herrchen].
Hier gibt es zwei Nominalphrasen als Kandidaten für einen Eigennamen (NE). “Der Struppi” bezeichnet eine einzige Einheit. Es kann auch mehrere Struppis geben, aber diese haben an sich keine gemeinsamen Eigenschaften, bis auf den gemeinsamen Namen, daher handelt es sich um einen Eigennamen. *”seinem Herrchen”* bezeichnet zwar (typischerweise) auch nur eine einzige Person allerdings können wir diese nur über die Eigenschaft identifizieren, dass sie ein Herrchen ist und dass dies für Struppi zutrifft. Struppi

könnte auch mehrere Herrchen haben, die alle die Eigenschaften teilen, die ein Struppi-Herrchen beinhaltet (z.B. darf Struppi streicheln, muss ihn ausführen und füttern etc.)
Schritt 3: Determinierer sind keine Teile des Namens.
Beispiel: *Der [Struppi]NE folgt seinem Herrchen.*
Schritt 4: Eigennamen können mehr als ein Token beinhalten. Beispiel:
Viele Personennamen (PER für person):
[Carl Linnaeus]PER
Buchtitle (OTH für other):
[Systema Naturae]OTH
Schritt 5: Eigennamen können auch in einander verschachtelt sein. Beispiel:
Personennamen in Filmtiteln:
[[Shakespeare]PER in Love]OTH
Orte (LOC für location) in Vereinsnamen (ORG für organisation):
[Hebarium Senckenbergianum [Frankfurt]LOC]ORG
Schritt 6: Titel, Anreden und Besitzer gehören NICHT zu einem komplexen Eigennamen. Besitzer können natürlich selber Eigennamen sein. Beispiel:
Referenz auf Musiktitel:
[Vivaldis]PER [Vier Jahreszeiten]OTH
Referenz auf Personen:
Landesvorsitzende Frau Vorstandsvorsitzende Dr. [Ute Wedemeier]PER
Schritt 7: Wenn das Gesamttoken einen Eigennamen darstellt, dann wird dieser annotiert. Beispiel:
Stiftungen: *[[Böll]PER-Stiftung]ORG*
Schritt 8: Kann in einem Kontext nicht entschieden werden, ob eine NP sich als Eigennamen oder Appellativ verhält, wird es nicht als NE markiert. Beispiel:
Ortsnamen vs. -beschreibungen:
...und zogen mit ihren grossen Transparenten gestern vom [Steintor] über den [Ostertorsteinweg]LOC zum [Marktplatz].
Schritt 9: Wenn ein Name als Bezeichnung für bestimmte Gegenstände in die Sprache übergegangen ist und in seiner Nutzung nicht als NE fungiert, so wird dieser nicht annotiert. Beispiel:
[Teddybär] (NICHT PER)
[Colt] (NICHT PER)
Schritt 10: Bei Aufzählungen mit Hilfe von Bindestrichen oder Vertragen eines Teils der NE auf spätere Wörter, wird die NE so annotiert, als sei sie voll ausgeschrieben.
Beispiel:
[Frühe]OTH und [Späte Bronzezeit]OTH
[Süd-]LOC und [Nordafrika]LOC

B.3 Zu welcher semantischen Klasse gehört ein Eigenname?

Wenn der Eigenname in eine der Klassen in der Liste Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens gehört, dann annotiere die zugehörige Klasse. Sollte die gefundene NE Rechtschreibfehler enthalten, wird sie dennoch annotiert. In Zweifelsfällen hilft auch die Tabelle NoSta-D-BIOfid-TagSet und alle Untertabellen, insbesondere die Beispiele mit dem weiter.
Jahreszahlen in ORGanisationen werden nicht markiert.
Beispiel:
[ICEI]ORG [2018]TIME
[Fussball-WM]ORG [2014]TIME
Wenn der Eigennamen in KEINE der vorhandenen Klassen passt, markiere diesen mit ***UNCLEAR***, notiere dir bitte das Beispiel und schicke uns eine E-Mail an: a.b@c.de. So können wir die Guidelines sukzessiv verbessern.

B.4 Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens:

- Elemente der fraglichen Einheit verbinden die gleichen Eigenschaften → Klasse → keine NE
- Christen glauben an Christus → Christ glaubt an Christus → keine NE
- Die Elemente der fraglichen Einheit verbindet nur der Name oder Element ist Einheit bezeichnet ein spezifisches Individuum → Name → NE
- *"Paleocene"* bezeichnet spezifische Epoche → NE (OTH)

NoSta-D-BIOfid-Tagset

Subcategory	Examples
person	*Carl Linnaeus*
Surname	*Tüxen, Tx.*
Artist names	*Madonna*
Charactere	*Schneewitchen, Miss Piggy*
Nicknames	*Sternchen333*
Superheroes	*Batman*

Table 6: Category 'PER-Person'

Subcategory	Examples
Hybrids	*Abies alba x Abies normannia*
Variety	*Asplenium scolopendrium var. crispum*
Form	*Araschnia levana f. prorsa*
Subspecies	*Falco peregrinus subsp. calidus, Pollichia semirubella ssp. semirubella*
Species	*Coeloglossum viride, Grüne Hohlzunge*
Genus	*Dendrocopus, Buntspechte*
Subfamily	*Phyticinae*
Family	*Noctuidae, Rosaceae*
Order	*Lepidoptera*
Class	*Aves, Insecta*
Phylum	*Chordata, Tracheophyta*
Kingdom	*Animalia, Plantae*

Table 7: Category 'TAXON': scientific and vernacular names (vernacular names only when referring to a certain taxon)

Subcategory	Examples
Districts	*Schöneberg*
Sights, Churches	*Brandenburger Tor, Johanniskirche*
Planets	*Mars*
Landscapes	*Königsheide*
Streets, places	*Söogestrasse, Alexanderplatz, A5*
Shopping centres	*Luisencenter, Allee-Center*
Mountains, lakes, rivers	*Alpen, Viktoriasee, Spree*
Continents	*Europa, Asien*
Countries, states	*Frankreich, Hessen, Assyrien, USA*
Cities	*Berlin, Babylon*
Regions	*Gazastreifen*

Table 8: Category 'LOC-Location'

Subcategory	Examples
Day	*Freitag*
Month	*Februar*
Year	*1835*
dd.mm.yyyy	*13.02.1835*
Century	*19. Jahrhundert*

Table 9: Category 'TIME'

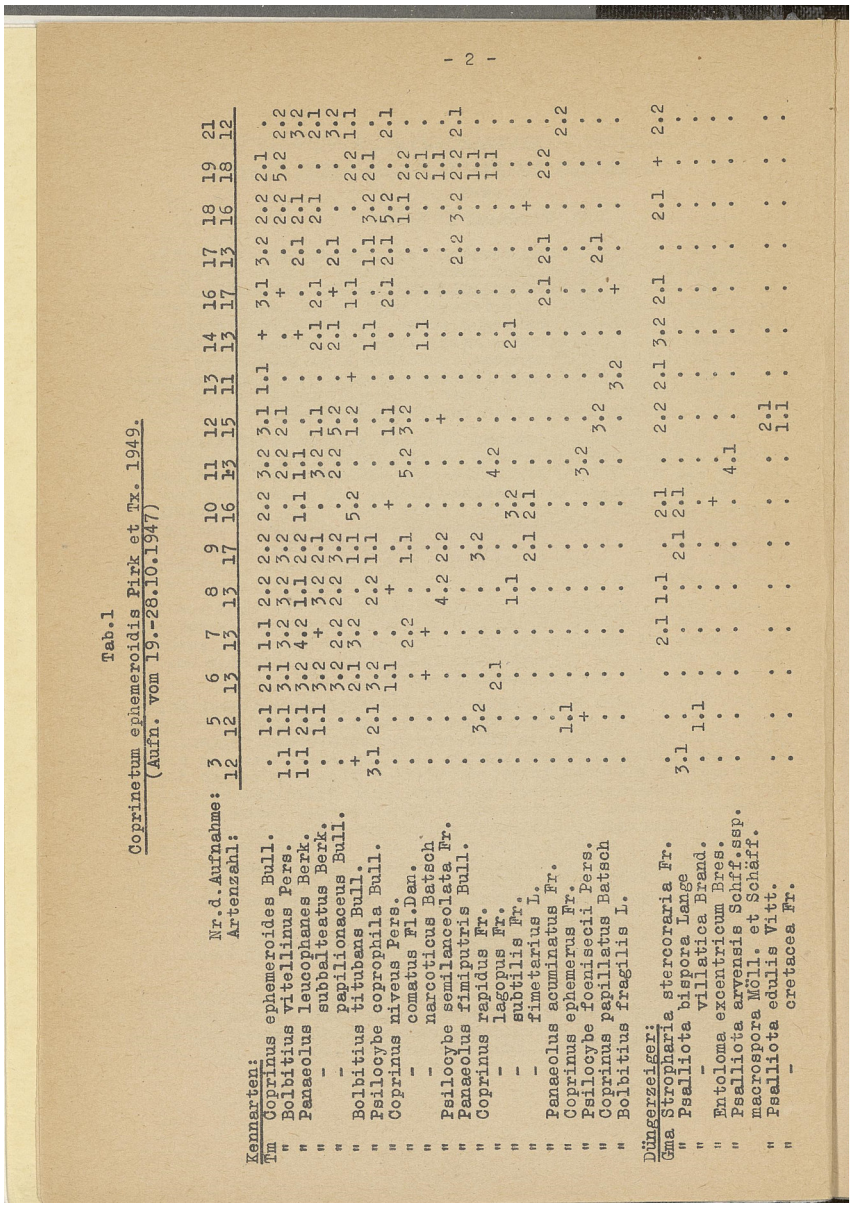
Subcategory	Examples
Book-, Film titles etc.	*Faust, Canon Medicinae*
Currencies	*Euro, Deutsche Mark*
Languages	*Deutsch, Latein*
Epochs	*Paleocene, Neolithikum, (auch Neubildungen: 'Neuzeit')*

Table 10: Category 'OTH-Others'

Subcategory	Examples
Organisations	*BHL, EU, Landgericht Frankfurt, Deutsche Botanische Gesellschaft*
Companies	*Microsoft, Bertelsmann*
Airports	*Fraport*
Operators	*Lotto 6 aus 49*
Institute	*Institut für Informatik*
Museums	*Senckenberg Museum*
Newspapers, journals	*Süddeutsche Zeitung, Nature, Beiträge zur Entomologie*
Clubs	*Eintracht Frankfurt*
Theatres, cinemas	*Metropol-Theater, CinemaxX*
Festivals	*Eurovision Song Contest, Berlinale*
Expositions	*Faszination Vielfalt*
Universities	*Goethe Universität Frankfurt*
Radio stations	*Arte, Planet Radio*
Restaurants and hotels	*Sassella, Mariott*
Military units	*Blauhelme*
Hospitals, Nursing home	*Charit, Klinikum Ingolstadt*
Fashion brands	*Chanel*
Sporting events	*Olympische Spiele, Wimbledon*
Bands	*Beatles, Die Fantastischen Vier*
Institutions	*DFG, Vogelwarte Helgoland*
Libraries	*UB J.C. Senckenberg*
Parties	*SPD, CDU*

Table 11: Category 'ORG-Organisation'

C Sample Page



Tab.1

Coprinetum ephemeroidis Pirk et Tx. 1949.

(Aufn. vom 19.-28.10.1947)

Nr.d.Aufnahme:	3	5	6	7	8	9	10	11	12	13	14	16	17	18	19	21
Artenzahl:	12	12	13	13	13	17	16	13	15	11	13	17	13	16	18	12
Kennarten:																
Tm Coprinus ephemeroides Bull.	.	1.1	2.1	1.1	2.2	2.2	2.2	3.2	3.1	1.1	+	3.1	3.2	2.2	2.1	.
" Bolbitius vitellinus Pers.	1.1	1.1	3.1	3.2	3.2	3.2	.	2.2	2.1	.	.	+	.	2.2	5.2	2.2
" Panaeolus leucophanes Berk.	1.1	2.1	3.2	4.2	1.1	2.2	1.1	1.1	.	.	+	.	2.1	2.1	.	3.2
" - subbalteatus Berk.	.	1.1	3.2	+	3.2	2.1	.	3.2	1.1	.	2.1	2.1	.	2.1	.	2.1
" - papilionaceus Bull.	.	.	3.2	2.2	2.2	3.2	.	2.2	5.2	.	2.1	+	2.1	.	.	3.2
" Bolbitius titubans Bull.	+	.	2.1	3.2	.	1.1	5.2	.	1.2	+	.	1.1	.	.	2.2	1.1
" Psilocybe coprophila Bull.	3.1	2.1	3.2	.	2.2	1.1	.	.	1.1	.	1.1	.	2.1	1.1	2.1	.
" Coprinus niveus Pers.	.	.	1.1	.	+	.	+	.	5.2	3.2	.	.	2.1	5.2	.	2.1
" - comatus Fl.Dan.	.	.	.	2.2	.	1.1	.	5.2	3.2	.	1.1	.	.	1.1	2.2	.
" - narcoticus Batsch	.	.	.	+	+	.	.	.	.	+	.	.	.	.	2.1	.
" Psilocybe semilanceolata Fr.	.	.	.	.	.	4.2	2.2	.	.	.	.	.	.	2.2	3.2	2.2
" Panaeolus fimiputris Bull.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1.1	2.1
" Coprinus rapidus Fr.	.	3.2	.	.	.	3.2	.	.	.	.	.	.	.	.	1.1	.
" - lagopus Fr.	.	.	2.1	.	.	.	.	4.2	.	.	2.1	.	.	.	1.1	.
" - subtilis Fr.	.	.	.	.	1.1	.	3.2	.	.	.	.	.	.	.	.	.
" - fimetarius L.	.	.	.	.	.	2.1	2.1	.	.	.	.	2.1	2.1	.	2.2	.
" Panaeolus acuminatus Fr.	.	1.1	.	.	.	.	.	.	.	.	.	.	.	.	.	2.2
" Coprinus ephemerus Fr.	.	+	.	.	.	.	.	3.2	.	.	.	.	.	.	.	.
" Psilocybe foeniseci Pers.	.	.	.	.	.	.	.	.	3.2	.	.	.	2.1	.	.	.
" Coprinus papillatus Batsch	.	.	.	.	.	.	.	.	.	3.2	.	+	.	.	.	.
" Bolbitius fragilis L.	.	.	.	.	.	.	.	.	.	.	3.2	.	.	.	.	.
Düngerzeiger:																
Gma Stropharia stercoraria Fr.	.	.	.	2.1	1.1	.	2.1	.	2.2	2.1	3.2	2.1	.	2.1	+	2.2
" Psalliota bispora Lange	3.1	.	.	.	.	2.1	2.1	.	.	.	.	.	.	.	.	.
" - villatica Brand.	.	1.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.
" Entoloma excentricum Bres.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.
" Psalliota arvensis Schff.ssp.	.	.	.	.	.	.	.	4.1	.	.	.	.	.	.	.	.
macrospora MÖll. et Schaff.																
" Psalliota edulis Vitt.	.	.	.	.	.	.	.	.	2.1	.	.	.	.	.	.	.
" - cretacea Fr.	.	.	.	.	.	.	.	.	1.1	.	.	.	.	.	.	.

Figure 3: Sample page from the digitized BIOfid corpus (taken from http://vl.ub.uni-frankfurt.de/biodiv/periodical/pageview/9028548).