ANNOTATIONSGUIDELINES ZUR LEMMATISIERUNG

A. Grundsätzliches:

Die Lemmatisierung findet mithilfe des Tools CAB (Cascaded Analysis Broker), das vom DTA (Deutsches Text Archiv) entwickelt wurde, statt. Diese automatische Lemmatisierung wird manuell kontrolliert und korrigiert. Dabei werden die Lemmata an die Eintragungen, die unter dem Layer Norm zu finden sind, angepasst. Die Lemmatisierung erfolgt als freie Annotation unter dem Layer Lemma.

B. Regeln:

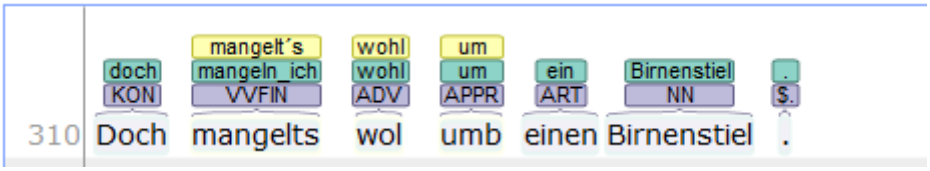
1.	Das Lemma orientiert sich immer an der Eintragung auf dem Layer „Norm“ oder an dem Token im Text, sofern keine Normkorrektur erfolgt ist, und gehört der gleichen Wortart an wie das Token.
2.	Die Lemmata werden dem jeweiligen Token zugeordnet. Die Annotation in diesem Layer erfolgt tokengenau. Dies bedeutet, dass bei zusammengesetzten Wörtern wie Schau-Platz die Lemmatisierung: „Schau“ „-“ „Platz“ ist. Ausnahmen bilden dabei: 1) Token, die keine Komposita darstellen und aus technischen Gründen nicht als ganzes Wort, sondern in einzelnen Wortfragmenten erfasst wurden (z. B. „ma“ + „chen“). In solchen Fällen wird dem ersten Fragment der Tag „OA“ zugeordnet (bei mehr als zwei Fragmenten auch den jeweils darauffolgenden) und dem letzten Fragment wird das eigentliche Lemma zugewiesen (z. B. „ma“ + „chen“ → „OA“ + „machen“). 2) Token(-fragmente), die nicht im Originaltext vorkommen und aus technischen Gründen falsch erfasst wurden. Nach einem Abgleich mit dem Originaltext wird diesen Token ebenfalls der Tag „OA“ zugeordnet. *OA = ohne Annotat (vgl. Barteld et al. 2016), im PoS-Layer nach STTS als XY zu annotieren
3.	Für die einzelnen Wortarten gelten folgende Regeln: • Nomen → Nominativ Singular *dazu zählen jegliche adjektivische, verbale u. a. substantivierte Formen • Verben → Infinitiv Präsens *Partizipien II als Teil des komplexen Verbs (ist *gelaufen*; hat *gelacht*) → Infinitiv Präsens • Adjektive → Positiv im Nominativ Singular *adjektivisch gebrauchte Partizipien I und II (*singender*; *gekochte*) → Kurzform des Partizips wie *singend* bzw. *gekocht* *Kardinalzahlen (*ein, eins, zweier*) → Nominativ Singular wie *eins* bzw. *zwei*

	*Ordinalzahlen (*erster, dritte, zehnter*) → Nominativ Singular wie *erst* bzw. *dritt* bzw. *zehnt* • bestimmte Artikel (*der, die, das…*) → immer *der* • unbestimmte Artikel (*ein, einer, eine, eines…*) → immer *ein* • Indefinitartikel (=indefinites Begleiterpronomen wie *jeder, jede, jedes; beider, beiden*; *jedermann, jedermanns*; *niemands*…) → erhalten nach Möglichkeit eine *er*-Endung wie *jeder* bzw. *beider* oder behalten die Form im Nominativ Singular wie *jedermann* bzw. *niemand* • Indefinitpronomen (=indefinites Stellvertreterpronomen wie *keiner, keine*; *beides, beidem, beide*; *einer, eine*; *jemandem*…) → erhalten nach Möglichkeit eine *e*-Endung wie *keine* bzw. *beide* bzw. *eine* oder behalten die Form im Nominativ Singular wie *jemand* *nicht flektierbare Indefinitpronomen bleiben so wie sie sind (*etwas*…) *Pronomen *man* bleibt so wie es ist, die Suppletivformen im Dativ und Akkusativ (*einem* bzw. *einen*) werden zu *eine* lemmatisiert • Personalpronomen (*ich, du* / *dich* / *dir, es, wir* / *uns, sie, Sie*) → immer *ich* • Reflexivpronomen bzw. reflexiv gelesene Pronomen (*mich* / *mir, dich* / *dir, sich, uns, euch*) → immer *mich* *nicht flektierbares Reflexivpronomen *einander* bleibt so wie es ist *für Unterscheidung von ‚regulär' gebrauchten Personalpronomen stärker auf Kontext achten • Possessivartikel (=possessives Begleiterpronomen wie *meines, unseren, Ihre; deinige, Eurigen…*) → immer 1. Pers. Sg., die eine *er*-Endung wie *meiner* bzw. *meiniger* erhält • Possessivpronomen (=possessives Stellvertreterpronomen wie *meiner, deinen*; *meinigen, Eurige*…) → immer 1. Pers. Sg., die eine *e*-Endung wie *meine* bzw. *meinige* erhält • Demonstrativartikel (=demonstratives Begleiterpronomen wie *dieser, diesem, diese; derjenige, diejenige*…) → erhalten nach Möglichkeit eine *er*-Endung wie *dieser* oder behalten die Form im Nominativ (Singular/Plural/(Mask.)) wie *derjenige* • Demonstrativpronomen (=demonstratives Stellvertreterpronomen *dieser, diesem, diese, derjenige, diejenige*…) → erhalten nach Möglichkeit eine *e*-Endung wie *diese* oder behalten die Form im Nominativ (Singular/Plural/(Fem.)) wie *diejenige* • Relativartikel (=relatives Begleiterpronomen wie *dessen, deren*) → immer *der* • Relativpronomen (=relatives Stellvertreterpronomen wie *der, die; welche, welches; wer, was*…) → *die* bzw. *welche* bzw. *wer* **was* wird als PRELS zu *wer* • Interrogativartikel (=interrogatives Begleiterpronomen wie *wessen, welcher*…) → wie Interrogativpronomen • Interrogativpronomen (=interrogatives Stellvertreterpronomen wie *wer, wessen, wem, wen*; *welcher, welche, welches*…) → *wer* bzw. *welcher* *Interrogativpronomen *was* bleibt so wie es ist *Interrogativadverbien (*wo, warum, wie*) bleiben so wie sie sind

	• Worte, die nicht flektiert werden können, bleiben wie sie sind. • Bei fremdsprachlichen Token wird dieses als Lemma übernommen. • Arabische und römische Ziffern → immer die arabischen Ziffern • Bei einer automatisch generierten Tokenkombination aus einem Satzzeichen (i. d. R. Anführungszeichen "..." / <<…>> / >…< oder Klammern […]) und einem Lexem (wie "Der / [Zweite), die INCEpTION als ein Token behandelt, werden nur die Lexeme lemmatisiert.

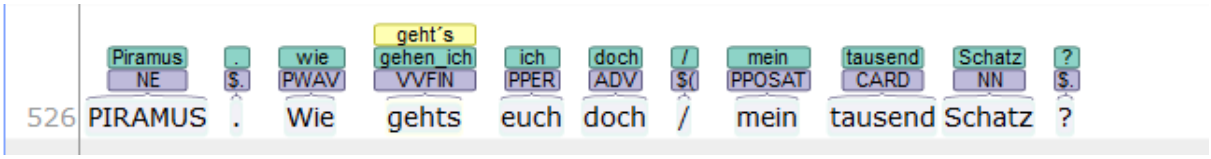
Beispiele zu ausgewählten Phänomenen

Beispiel 1: Squentz



Text-Ebene: mangelts
Norm-Ebene (gelb): mangelt´s
Lemma-Ebene (grün): mangeln_ich

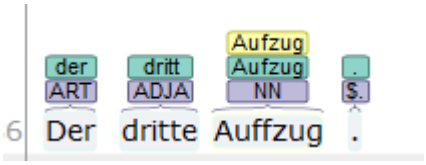
Beispiel 2: Squentz



Textebene: gehts
Norm-Ebene (gelb): geht´s
Lemma-Ebene (grün): gehen_ich

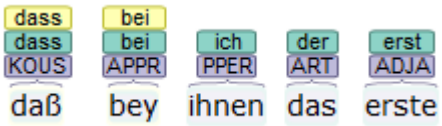
Ordinalzahlen (*erster, dritte, zehnter*) werden wie Adjektive behandelt und somit auch eine nicht flektierte Form zurückgeführt.

Beispiel 3: Squentz

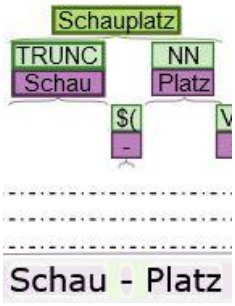


Text-Ebene: dritte
Norm-Ebene (gelb): keine Annotation notwendig
Lemma-Ebene (grün): dritt

Beispiel 4: Squentz

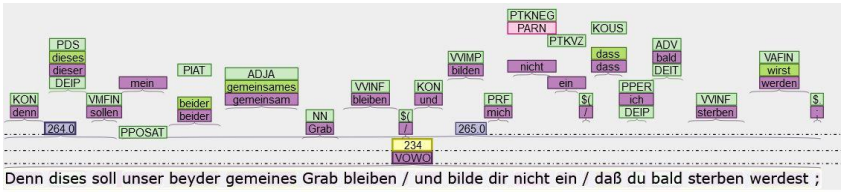


Text-Ebene: erste
Norm-Ebene (gelb): keine Annotation notwendig
Lemma-Ebene (grün): erst
Beispiel 5: Cardenio und Celinde



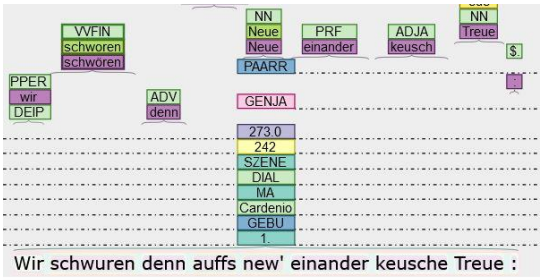
Text-Ebene: Schau - Platz
Norm-Ebene (hellgrün): Schauplatz
Lemma-Ebene (mintgrün): Schau – Platz

Beispiel 6: Cardenio und Celinde



Text-Ebene: dir
Norm-Ebene (hellgrün): keine Annotation notwendig
Lemma-Ebene (mintgrün): mich

Beispiel 7: Cardenio und Celinde



Text-Ebene: einander
Norm-Ebene (hellgrün): keine Annotation notwendig
Lemma-Ebene (mintgrün): einander

C. Technisches Vorgehen:

1.	Einloggen in INCEpTION mit Benutzerkennung.
2.	*Gryphius-Projekt* bzw. *Pronomen in Dramen-Projekt* auswählen und auf *Annotation* klicken.
3.	Das Dokument öffnen, für dessen Korrektur/Annotation man eingeteilt wurde (Dokument anklicken).
4.	Gegebenenfalls unter *Settings* (Zahnrad-Symbol) für den Layer *Lemma* ein Häkchen setzen. In der Regel werden alle Layer automatisch ausgewählt.
5.	An der rechten Seite den Layer *Lemma* auswählen.
6.	Das jeweilige Token für die Annotation auswählen, indem das über dem Token liegende *Lemma*-Kästchen angeklickt wird. Auf der rechten Seite erscheint das Annotationsfeld.
7.	Korrektur vornehmen und mit Enter bestätigen.
8.	Gegebenenfalls auf *Clear* klicken, um den Layer zu schließen.

Die vorliegende Version ist eine Anpassung der Grundlagenguidelines aus: Eggert, Lisa; Müller, Melissa (2021): „LEMMA – Lemmatisierung“. In: „Guidelines. Interaktionale Sprache bei Gryphius“ URL: https://gryphiusprojekt.wordpress.com/lemma/