EmpiriST2015:
Ergänzungsdokument zu den Annotationsrichtlinien

Dieses Dokument gibt Hilfestellungen zu Problemfällen und “schwierigen” Kategorien beim PoS-Tagging und bei der Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation im Rahmen der Shared Task EmpiriST2015.
Grundlegend und maßgeblich für die Annotation sind die Richtlinien, die in den offiziellen Richtliniendokumenten zur Tokenisierung und zum PoS-Tagging dargestellt sind (https://sites.google.com/site/empirist2015/home/annotation-guidelines). Das vorliegende Dokument hebt diese Richtlinien nicht auf, sondern gibt zu einigen der verwendeten PoS-Kategorien sowie zu einzelnen Problemfällen weitergehende Präzisierungen und Beispiele.
Das Dokument wird vom Vorbereitungsteam der Shared Task ggf. sukzessive zum Projekt weiter ergänzt. Für Fragen zum Projekt und zu den Richtlinien gibt es die GoogleGroup https://groups.google.com/d/forum/empirist2015.

Letzte Bearbeitung: 5. Dezember 2015

Table of Contents:

Abgrenzung von Interjektionen (ITJ) zu anderen PoS-Klassen

- **BEISPIELE:**

  “echt ?” (Trial: social_chat.txt, posting 1-13)

  “was echt zori ?” (posting 1-3)

  → **Richtlinie:** Wörter, die formgleich auch in anderen Wortartenklassen (z.B. ADJ) vorkommen, werten wir nur dann als ITJ, wenn sie (a) nicht syntaktisch integriert sind und (b) die Äußerung keine propositionale Lesart zulässt. “echt” in den beiden Beispielen wird daher als ADJD getaggt, da eine propositionale Lesart angenommen kann (i.S.v. “Ist das wirklich wahr?” oder “Das glaube ich nicht!”).
- **Charakteristisch für Interjektionen (vgl. z.B. *GRAMMIS*):** nicht syntaktisch integriert; keine propositionale Lesart; Funktion im Bereich der Interaktionsorganisation oder emotionalen Kommentierung.

Formen von “okay” (*okay/OK/O.K.*): Tokenisierung und PoS-Klassifikation (PTKANT, ADJD)

Tokenisierung:

- Abschnitt 4.1 der Tokenisierungs-Richtlinie sieht vor, dass mehrgliedrige Abkürzungen in einfache Abkürzungen aufgetrennt werden (also z.B. <d.h.> -> <d.> <h.>). Für Formen von “okay” mit Abkürzungspunkten (“O.K.”, “o.k.”) gilt die folgende Ausnahme: **Formen von “okay” (also auch “O.K.” und “o.k.”) werden unabhängig von ihrer Schreibweise grundsätzlich als *ein Token* behandelt.** Grund: Die mehrgliedrige Abkürzung ist im Deutschen nicht mehr transparent.

PoS-Klassifikation:

- OK und okay werden nach ihrer syntaktischen Funktion entweder als PTKANT oder als ADJD getaggt.

  Richtlinie:
- PTKANT liegt vor, wenn die Einheit syntaktisch nicht integriert ist und responsiv verwendet ist. Das ist in nachfolgend (1), (2) und (3) der Fall. Im (1) ist die PTKANT mit einer ITJ (“na!”) kombiniert.
- ADJD liegt vor, wenn die Einheit als (adjektivische) Ergänzung im Rahmen einer Kopulakonstruktion fungiert. Das ist in nachfolgend (4) und (5) der Fall.

Beispiele:

- **(1) „na okay“**

  (2) „ok, dann schau ich ma eben“

(3) „O.k. dann schreibe ich bis morgen drauf“

(4) „wäre auch ok“ / “ist o.k.!”

(5) „ist es ok, wenn unser geplänkel drin stehen bleibt?“

Behandlung der Form “Re” in Chats (ITJ)

Richtlinie: Formen von “re” werden, wenn sie wie eine Grußformel verwendet sind, als ITJ getaggt. “Re” bedeutet in diesem Fall so viel wie “Wieder hallo”.

Verwendungsbeispiele:

- ***mieze:*** **re Happy**

  (Begrüßung des Chatters Happy durch Chatterin mieze, nachdem Happy vorübergehend den Chat-Raum verlassen hat und wieder zurückkehrt.)
- ***Petra:*** **re :-)**

  xyz: „**Reeeee**!

  (Chatterin Petra begrüßt bei ihrem Wieder-Eintritt in den Chat-Raum die anwesenden anderen Chatter mit “re”. Chatter xyz erwidert den Gruß mit “Reeeee”; die Funktion ist, ungeachtet der Graphemiteration, dieselbe.)

Zuweisung der neuen Partikelklassen PTKIFG, PTKMA und PTKMWL

Die Präzisierung zu den nachgenannten Fällen erfolgt in Übereinstimmung mit den Guidelines zur PoS-Annotation von Daten gesprochener Sprache im Projekt “Forschungs- und Lehrkorpus Gesprochenes Deutsch” (FOLK) am Institut für Deutsche Sprache, Mannheim) (IDS).

PTKIFG (Intensitäts-, Fokus- und Gradpartikeln)

- Lassen sich bei einer Umstellprobe nur mitsamt ihrer Mutterphrase vor das finite Verb stellen: “der Joghurt ist **voll gut**” → “**voll gut** ist der Joghurt”
- Adverbien (ADV) lassen sich im Gegensatz zu PTKIFG auch alleine vor das finite Verb stellen: “der Joghurt ist **noch** gut” → “**noch** ist der Joghurt gut”
- PTKIFG, die von Adjektiven abgeleitet sind, lassen sich ebenfalls durch Umstellprobe von der Kategorie ADJA abgrenzen.
- Weitere Beispiele: “das stimmt **gar** nicht”, “das ist **voll** schön”, “**gerade** du musst das sagen”, “du bist echt **selten** doof”

PTKMA (Modal- und Abtönungspartikeln)

- PTKMA grenzen sich von PTKIFG distributionell dadurch ab, dass sie sich im Satz typischerweise nicht umstellen lassen.
- Beispiele: “das ist **halt** Pflicht”, “was gehst du **auch** immer so spät ins Bett?”, “jetzt warte **mal**”, “wie kann man sich **nur** so was Gefälschtes kaufen?”, “kriegen wir **schon** irgendwie hin”

PTKMWL (Partikeln als Teile von Mehrwortlexemen)

- drücken gemeinsam oft Aspekt aus; “Aspektpartikeln” (Hardarik Blühdorn).
- Lässt man einen Teil des Mehrwortlexems weg, verändert oder verliert es seine Bedeutung.
- Bei einer Umstellprobe lassen sich nur beide Teile des MWL gemeinsam in das Vorfeld stellen.
- Beispiele: “Ich sehe **nichts mehr**”, “er kommt **immer noch** zu spät”, “wir haben das **gerade erst** gesehen”, “da ist er **schon wieder**” (die PTKMWL ist jeweils rot hervorgehoben).

Adverbien (ADV) in Verbindung mit PTKMWL

- Wenn Adverbien zusammen mit Partikeln ein Mehrwortlexem bilden, ist folgende Besonderheit zu berücksichtigen: Während für ADV typischerweise gilt, dass sie alleine ins Vorfeld verschiebbar sind, so ist dies für ADV in Mehrwortlexemen (z.B. für “immer” in “schon immer”) gerade nicht der Fall: Mehrwortlexeme können nur als Ganze ins Vorfeld gerückt werden.

Best practices zur Behandlung einzelner Wörter, die nur in bestimmten Verwendungen als Partikeln fungieren (*auch*, *Mal/mal*)

(In anderen Fällen kann analog bzw. unter Anwendung ähnlicher Operationen und Tests entschieden werden.)

auch

- **ADV:** Bezug auf Verb; *auch* kann als Konjunktionaladverb im Vorfeld sowie in Kombination mit kognitiven Verben auftreten.

  ⇒ Im Vorfeld: “**Auch** hatte niemand daran gedacht…”;

  ⇒ Im Mittelfeld: “Ich meine **auch**, dass…”
- **PTKIFG:** Bezug auf eine bestimmte Phrase; hierbei drückt *auch* zusammen mit seinem Bezugsausdruck eine Alternative bzw. einen weiteren Faktor des Gesagten

aus. Die Bezugsausdrücke können beliebig komplex sein, *auch* steht dabei meistens vor dem Bezugsausdruck, kann aber auch dahinter bzw. in Distanzstellung stehen, jedoch nie alleine im Vorfeld.

⇒ Im Vorfeld mit Konjunktion: "**auch** wenn wir diese Sache schon besprochen haben…"

⇒ Im Vorfeld mit Bezugsausdruck: "**Auch** der Leo hat ne Sonnenbrille"; "**Auch** im Detail muss man einfach vieles hinterfragen"

⇒ Im Mittelfeld: "ja, das ist **auch** eine Schlussfolgerung"; "die Folien sind **auch** wirklich gut"; "ich habe **auch** gearbeitet heute"

- **PTKMA:** Hier ist *auch* an das Mittelfeld gebunden, bildet keine Phrase und kann nicht erfragt werden. Wie alle anderen Modalpartikeln kann *auch* mit anderen Partikeln, wie z.B. *ja* kombiniert werden und tritt häufig in Frage- oder Aufforderungssätzen auf. Bsp.: "Warum gehst du **auch** immer so spät ins Bett?"; "Wie **auch** immer", "du musst ja **auch** immer petzen"

Mal, mal

- **NN:** Verwendung als Nomen; Bsp.: "das erste Mal"
- **ADV:**
  - als umgangssprachliche Kurzform für das temporale Adverb „einmal" (Test: ersetzbar durch „irgendwann", „über kurz oder lang", „ab und zu"). Drückt aus, dass eine Handlung nicht sofort, sowie nicht dauerhaft stattfindet, insbesondere erstere Bedeutung unterscheidet das Adverb von der Modalpartikel! Bsp.: "**mal** kann man das machen"
  - Verwendung als nicht nachfeldfähiger Adverbkonnektor: "**mal** so, **mal** so"
  - Verwendung als umgangssprachliche Kurzform als Teil eines Mehrwort-Lexems, mit dem zusammen es meist Aspekt ausdrückt, d.h. PTKMWL+ADV: "schon **mal**", "noch **mal**", "erst **mal**"
- **PTKMA:**

  a) Allgemein: Nicht erfragbar; lässt sich nicht eins zu eins ins Englische übersetzen bzw. ist für die Übersetzung meist irrelevant; meist Bezug auf aktuelle Situation, insb. Sätze mit Aufforderungscharakter (abmildernde Wirkung)

  b) Häufige Verwendungen:

  - als Teil von Imperativen, Aufforderungen: "pass **mal** auf"; "guck mal"; "sag **mal**"; "bleiben Sie **mal** bei dem was Ihnen ihr Bauch sagt";

- Sprecher stellt Hypothesen auf: “ich sag **mal**”; “nehmen wir **mal** an”;
- zusammen mit temporalen Ausdrücken, aber auch syntaktischen Konstruktionen und Kontext, die eine adverbiale Verwendung von *mal* unmöglich/unnötig machen: “dann habt ihr jetzt halt **mal** nichts zu tun”; “heute **mal** nicht”; “ich spiele ab und zu **mal** sehr gerne Gitarre”.

noch

- **ADV:**

  a) Wenn *noch* vor das finite Verb gestellt werden kann und für eine adverbiale Bestimmung der Zeit steht: “Ich habe die **noch** nicht reingetan”, “ich warte **noch**”

  b) Bei nicht eindeutiger Leseart bzw. ambigen Sätzen ist *noch* als ADV die präferierte Leseart: “Weiß nicht, wie viel ich **noch** hab”, “und unterhält sich **noch** mit der Kindergärtnerin”

- **PTKMWL:**

  a) In Verbindung mit Kopf-Lexem: “**Noch** X” bzw. “X **noch**”; treten immer zusammen mit einer NP oder einer Kardinalszahl auf. Im Unterschied zu einer Fokuspartikel markiert *noch* hier Aspekt: “auch **noch**”, “**noch** mal”, “**noch** etwas”, “**noch** so”, “immer noch“

  b) Bei Bezug auf Fragepronomen (im Sinne von “außerdem”): “Und wen **noch**?”, “Was haben die **noch** gespielt?”, “Welche anderen Werke können Sie denn **noch** nennen?”

- **KON:**

  ⇒ Wenn noch zu der nebenordnenden mehrteiligen Konjunktion *weder...noch* gehört: “Weder Äpfel **noch** Birnen”, “Sie konnten weder laufen **noch** kriechen”

nur

- **ADV:**

  ⇒ Vorfeldfähig, adversative Bedeutung, schränkt die Aussage der vorangegangenen Äußerung ein

  ⇒ Häufige Fälle:

  a) Konjunktionale Verwendung: “**Nur** ist da zehn Minuten Unterschied [...] gewesen”, “Das kann sein [...], **nur** ist das einfach auch mal anders”

  b) Hinter kognitiven Verben (*denken, meinen, glauben, wissen*…): “Ich dachte **nur**,

Sie fragen jetzt", "ich meinte **nur** falls der", "ich weiß **nur**, dass Lena in letzter Zeit immer von den Prüfungen redet", "ich sag **nur** Stichwort Leistungsfähigkeit"

- **PTKIFG:**

  a) Bezug auf Nomen und Nominalphrasen: "Weil des **nur** Druck erzeugt", "Schwarzmeer ist einfach **nur** ein Ort"

  b) Bezug auf Präpositionen und Präpositionalphrasen: "Nee **nur** bei der Mutter", "Wir fragen das **nur** zur Erläuterung"

  c) Bezug auf Adverbien: "Wenn der nich **nur** heut der ist", "Aber da ging des **nur** so"

  d) Bezug auf Adjektive und Adjektivphrasen: "Okay **nur** ganz kurz eben"

  e) Bezug auf Nebensätze: "**Nur** weil jetzt das Aufnahmegerät da liegt", "Auf dich ist immer Verlass, **nur** wenn ich was bei dir zahlen muss dann bis du nicht nett"

  f) Bezug auf Pronomen und Pronominalphrasen: "Ein bisschen **nur** oder wie", "nicht **nur** in dem Zeitpunkt anzusiedeln"

  g) Bezug auf Kardinalzahlen: "**Nur** zwei Zentimeter drunter"

  h) Bezug auf Partizipien: "Nee ich hatte des **nur** nachgeguckt und da hab ich **nur** gesehen okay die schließen das aus", "Die ist da aber wieder **nur** gedacht die Linie ne"

  i) Bezug auf Infinitive, meist in Kombination mit finitem Modalverb: "Des wollt ich **nur** noch mal bemerken", "Ja ich will ihn **nur** grad fragen was passiert ist", "Sodass die **nur** noch zumachen müssen", "Einfach **nur** festhalten net drücken oder so"

- **PTKMA:**

  ⇒ Drückt spezifische Sprechereinstellung aus: "Aber es geht auch mit Sprachen wenn man **nur** will", "Wie kann man sich **nur** so was Gefälschtes kaufen?", "Wir sind die vier besten Freunde die man sich **nur** wünschen kann", "**Nur** zu!", "**Nur** Mut!", "Was hast du **nur**?", "Wenn er **nur** käme"

- **DM:**

  ⇒ steht im Vor-Vorfeld: "Da fällt mir ein ich muss auch noch meine tollen Thrombosestrümpfe anziehen **nur** wo soll ich das machen"

schon

- **ADV:**

⇒ vorfeldfähig ohne Bedeutungsverschiebung

⇒ in Fragesätzen: schon kann ohne Bedeutungsverschiebung in der Antwort wiederholt werden

⇒ Beispiele: “sehr schön, ja die Funkstecke ist **schon** ionisiert”, “Sind die Kamele vielleicht **schon** draußen?”

- **PTKIFG:**

⇒ kann nur mit Bezugsphrase ohne Bedeutungsverschiebung ins Vorfeld verschoben werden

⇒ tritt vor oder (seltener) nach einer Bezugsphrase (meist einer Zeitangabe) auf

⇒ in V-1-Fragesätzen ohne temporale Bedeutung

⇒ Bedeutung: der in der Bezugsphrase genannte Zeitpunkt ist früher oder später als der erwartete, übliche Zeitpunkt, oder der in der Bezugs-Phrase genannte Wert ist größer als erwartet

⇒ Beispiele: “drei Jahre **schon**”, “**schon** um elf Uhr?”

- **PTKMA:**

⇒ nicht vorfeldfähig ohne Bedeutungsverschiebung

a) in Aussagesätzen ohne Zukunftsbezug: Einräumung oder Zustimmung in Bezug auf den Sachverhalt, der im *schon*-Satz dargestellt wird: "umdrehen **schon**, aber sonst nichts"

b) in Aussagesätzen mit Zukunftsbezug: drückt Zuversicht in Bezug auf den Sachverhalt, der im schon-Satz dargestellt wird, aus: “kriegen ma **schon** schon irgendwie hin”

c) In W-Fragen ohne temporale Bedeutung (rhetorische Frage): “Was weiß der **schon**?”, “Und wenn **schon**?”

d) In Imperativsätzen: “Mach **schon**!”, “Jetzt sag **schon**!”

- **PTKANT:**

⇒ nicht in einen Satz eingebunden

⇒ als Antwort auf einen V-1-Fragesatz oder als Reaktion auf eine Aussage des Gegenübers

⇒ Bedeutung: Zustimmung bzgl. der Aussage des Gegenübers, die aber gleich eingeschränkt werden soll (implizit oder explizit)

⇒ Beispiel: “(ja,) **schon**”

- **PTKMWL:**

  ⇒ kann nur mit Bezugsphrase ohne Bedeutungsverschiebung ins Vorfeld verschoben werden

  ⇒ Bezugswort ist kein Nomen oder Zahlwort

  ⇒ Beispiele: “**schon** mal”, “vorhin **schon**”, “gestern **schon**”, “**schon** heute”

wie

- **PWAV:**

  a) Interrogativpronomen:

  ⇒ nicht unbedingt in der Bedeutung "auf welche Art und Weise"

  ⇒ Beispiele: "**Wie** geht es dir?", "Herr Feig, **wie** sieht es aus?". "**Wie** lang sind unsere Brennspannungen noch mal?"

  b) Relativpronomen:

  ⇒ mit Verb-letzt Stellung

  ⇒ lässt sich als Frage umformulieren (bei gleicher Bedeutung)

  ⇒ lässt sich durch "auf welche Art und Weise" ersetzen (außer bei Kombinationen wie "wie viel", "wie lang", usw.)

  ⇒ Beispiele: "klar zu machen **wie** früher Ausbildung läuft und **wie** heute", "[...] **wie** der individuelle Zugang erfolgt", "viel zu arg **wie** die da rausfahren"

  ⇒ Problemfälle: "**wie** auch immer"

- **KOKOM:**

  a) Illustrativer Adjunktor:

  ⇒ durch *wie* werden ein oder mehrere NPs als Adjunkte angegliedert, die mit ihrem Bezugswort (in der Regel direkt vor *wie*) in Kasus übereinstimmen

  ⇒ dient dazu, das Bezugswort zusatzlich zu charakterisieren

  ⇒ Beispiele: "Grundfragen **wie** Liebe, Tod, [...]", "auf Sachen **wie** Wortstellung, [...]", "sowas **wie** X gefällt mir"

  b) Konnotierender Adjunktor:

  ⇒ die *wie*-Phrase bestimmt entweder das mit dem Bezugsausdruck Gesagte genauer: "Einen Arzt **wie** Dr. Klaus findet man nicht so leicht."

⇒ oder der Inhalt der *wie*-Phrase steht im Vordergrund und kann das Bezugswort ohne große Bedeutungsänderung ersetzen: "Einen Menschen **wie** ihn muss man einfach gern haben!"

c) NP-, ProP oder PP-Bezug (komparativ):

⇒ Vergleich: etwas (ist) (genauso) wie etwas anderes

⇒ ohne Satz!

⇒ Beispiele: "Das ist ja **wie** bei den Pfalzwerken", "**Wie** bei den Indianern", v "Denen geht es **wie** mir"

- **KOUS:**

a) Temporalsätze mit *wie*

⇒ *wie* als temporalen Nebensatz einleitende Konjunktion

⇒ nicht standardsprachliche Verwendung

⇒ meistens durch *als* oder *während* ersetzbar

⇒ Beispiele: "**Wie** du das sagst, fällt mir ein, [...]", "die Katze, die sonst losrennt, **wie** sie den Hund erblickt [...]", "**Wie** sie eintritt, klingelt das Telefon"

b) Gleichzeitigkeit des Wahrnehmens signalisierend

⇒ Einleitungselement eines finiten Objektsatzes nach Verb der geistig-sinnlichen Wahrnehmung

⇒ Bedeutet eher "Ich nahm wahr, dass etwas geschah" als "Ich nahm wahr, auf welche Art und Weise etwas geschah"

⇒ Beispiel: "Ich sah, **wie** du ergriffen wurdest."

c) redekommentierender *wie*-Satz

⇒ Beispiele: "**wie** Sie eben schon sagten", "durch Einsatz dieser Fragenkataloge, **wie** Sie das vorgeschlagen haben", "zum Beispiel, **wie** Kaspar gesagt hat, [...]", "**Wie** gesagt, Herr Schmidt war gestern hier."

- **KON:**

a) Einteiliger kopulativ-komparativer Konjunktor

⇒ durch *und* ersetzbar

⇒ Verknüpfen in Bezug auf ein Charakteristikum gleiche Konjunkte, die gleiche aber nicht gemeinsame Geltung bekommen

⇒ Beispiele: "Männer **wie** Frauen strömten in den Saal", "Die Grünen erzielten

hier **wie** dort achtbare Ergebnisse"

b) Paariger kopulativ-komparativer Konjunktor

⇒ sowohl… **wie** (auch)

⇒ Beispiele: "Ich kenne sowohl den Vater **wie** auch den Sohn", "sowohl väterlicherseits **wie** mütterlicherseits"

- **PTKIFG:**

⇒ "wie" leitet keinen Nebensatz ein

⇒ Adjektiv- oder Adverbbezug

⇒ exklamativ

⇒ lässt sich nicht durch "auf diese Art und Weise" ersetzen

⇒ Beispiele: "Ich find es geil da. **Wie** der Himmel blau ist", "**Wie** eklig, sagt Oma", "Krass **wie** schnell die da drauf reagieren"

Abgrenzung ADJD vs. VVPP

- **BEISPIEL: "wir sind alle erlöst und kommen zum Vater"** (trial008.txt, #146)
  → VVPP qua *Verlaufspassiv* (alle Passivkonstruktionen enthalten VVPP).
- **BEISPIEL: "das ist damit gemeint"** (ibid., #302)
  → ADJD qua *Kopulakonstruktion* (Ergänzungen im Rahmen von Kopulakonstruktionen sind grundsätzlich als ADJD zu taggen).
- **Vorschlag für das Vorgehen bei der Annotation:** möglichst nach Disambiguierungs-Kriterien aus den STTS-Guidelines zwischen Kopulakonstruktion und Verlaufspassiv unterscheiden; wenn damit nicht eindeutig entscheidbar, dann nach intuitiver Interpretation. In einer Kopulakonstruktion wird *immer* das Tag ADJD verwendet, in Passivkonstruktion immer VVPP. Zitat aus den STTS-Guidelines (1999):
  1. *Verdacht auf VVPP: kann der Satz ins Aktiv gesetzt werden mit gleicher Semantik? Ja → VVPP*
  2. *regiert "von"-PP oder ähnliche Konstruktion, die auf Verbsemantik hinweist → VVPP*
  3. *Ersetzung durch semantisch nahes Adjektiv (das nicht von VVPP abgeleitet ist) möglich → ADJD*

Abgrenzung ADJD vs. ADV vs. PTKVZ

- **BEISPIEL: “es liegt ja auch nahe”** (trial007.txt, #64)
  → eindeutig PTKVZ (“naheliegen”), da der Satz ohne “nahe” ungrammatisch würde (bzw. eine völlig andere Bedeutung erhielte).
- **BEISPIEL: “dass wir hoch auf die Aelggialp fahren”** (trial007.txt, #43)
  → eindeutig ADJD. Ein PTKVZ läge vor in Fällen wie "Ich fahre auf die Aegialp hoch". Im *dass*-Satz mit Verbletzt-Stellung müsste der PTKVZ immer mit dem Verbstamm zusammen in Endposition stehen ("dass wir auf die Aegialp hochgefahren sind"); vgl. analog die folgenden (als ungrammatisch zu wertenden) Fälle mit Partikelverben in Verbletztsätzen: **...dass ich an mit dem Schreiben gefangen habe*, **... dass ich weg den Müll geworfen habe.*

Behandlung des Akronyms “aka”

- **BEISPIEL: “The Left Foot of God aka Bronislaw Bitchinski”** (trial006.txt, #17)
  → wird als KOKOM behandelt

Abgrenzung NN vs. NE vs. FM

- **BEISPIELE:**
  “The Left Foot of God” (trial006.txt)
  “Electric Herryland Studios” (trial006.txt, #162–#164)
  → **Richtlinie hierzu:** Mehrteilige fremdsprachliche Eigennamen werden komplett als NE getaggt (*nicht* als FM); dabei wird jedes Token, das dem mehrteiligen Ausdruck angehört, als ein NE-Vorkommen behandelt. Im ersten Beispiel oben folgen somit fünf NEs aufeinander.
  Andere fremdsprachliche Ausdrücke und Zitate (z.B. “God save the Queen”, “à la carte” oder “dialog on demand”) werden als FM getaggt, sofern sie nicht als Fremdwörter in den deutschen Sprachschatz eingegangen sind.
- **BEISPIEL: “Boss RC-50 Loop Station”** (trial006.txt, #25–#29)
  → “Boss” als NE
  → “RC-50” als XY (kein NE, weil damit ein Modell und keine einzelne Entität bezeichnet wird)
  → “Loop Station” als NN NN, da als Fremdwort in den deutschen Sprachschatz eingegangen

- **BEISPIEL: “Singlenot-**(sic!) **Funkriff”** (trial006.txt, #119-#120)
  → TRUNC NN, da auch diese Wörter im Deutschen inzwischen gängig sind. Man beachte, dass das Kompositum “Singlenote-Funkriff” im Originaltext irrtümlich getrennt geschrieben ist, was gem. der Tokenisierungs-Guidline bei der Annotation nicht rückgängig gemacht wird.

CARD vs. ART im Falle von “ein/eine”

- **BEISPIEL: “wandert auschließlich**(sic!) **in eine Tasche”** (trial006.txt, #56)
  → **Richtlinie:** Wir behandeln Fälle von “ein/eine” standardmäßig als ART. Das Vorliegen einer CARD nehmen wir nur in Fällen an, in denen im Kontext weitere CARD auftreten, die eindeutig die Lesart ermöglichen, dass “ein” in diesem Fall zur Quantifizerung verwendet ist. Das ist z.B. bei *ein bis zwei Millionen* der Fall; in allen anderen Fällen, in denen auch eine *nur* determinierende Lesart möglich ist, nehmen wir ART an. Entsprechend ist “eine” in “eine Tasche” als ART zu behandeln, ebenso “eine” in *Peter hat im Lotto eine Million gewonnen*. Diese Richtlinie ist relativ strikt und blendet u.U. manche Verwendungen von “ein/eine” aus, die im Kontext u.U. eine Lesart als CARD zulassen; bei weniger strikter Richtlinie wäre aber zu erwarten, dass viele Fälle von unbestimmten Artikeln ebenfalls strittig werden (z.B. in “Ich habe eine Tante, die Maria heißt”).

Schnellschreibphänomene: Behandlung von unvollständigen bzw. nicht interpretierbaren Wortformen und Wortteilen

- **BEISPIEL: “mich nochmal zu hi n hinstell”** (Trial: social_chat.txt, posting 1-23)
  → In diesem Fall ist davon auszugehen, dass aufgrund geringer Planung und/oder fehlendern Monitorings bei der Beitragsproduktion (a) die Verbpartikel “hin” zweifach realisiert und (b) in die Wortform “hin” versehentlich ein Leerzeichen eingefügt wurde. Alternativ könnte man “hi n” auch als Versuch der Realisierung von “ihn” oder “ihm” deuten, in den sich mehrere Tippfehler eingeschlichen haben. Welcher Deutung man sich auch anschließt: Eine eindeutige Interpretation ist nicht möglich.
  Für Fälle wie diese legen wir fest, dass sämtliche als Tokens konstituierte, unvollständige bzw. nicht interpretierbare Wortformen und Wortteile als XY (“Nichtwörter”) behandelt werden. Im o.a. Beispiel ergibt sich somit das folgende Tagging: “hi/XY n/XY hinstell/AKW”.

Tokenisierung komplexer Eigennamen

- Eigennamen ohne Leerzeichen bilden grundsätzlich ein einzelnes Token und werden *nicht* segmentiert. Die Anweisung der Tokenisierungs-Guidelines "Ausdrücke, die aus einem Kurzwort (Akronym) und einer Zahl bestehen, werden in zwei Tokens zerlegt" (S. 15) bezieht sich nur auf Fälle wie "WS04", die eine Kontraktion aus zwei eigenständigen Wörtern ("Wintersemester", "2004") bilden.
- Dementsprechend "DRSSTC3", "stART12", … (die Zahl ist fester Bestandteil des Eigennamens, auch wenn sie im letzteren Fall strenggenommen für das Jahr "2012" steht).
- Nicht segmentiert wird auch "[gab_log]", der etwas ungewöhnliche Name eines Blogs für Geisteswissenschaftler. Dieser Fall ist in den Trial-Daten (trial_009.txt) falsch tokenisiert, da die Klammern irrtümlicherweise als Satzzeichen verstanden wurden.
- Entsprechend behandelt werden Eigennamen, die wie komplexe Abkürzungen aussehen: "B.Z." und "S.H.I.E.L.D." bleiben als ein Token erhalten. ("B.Z." steht nicht für die *Berliner Zeitung*; aber selbst wenn es so wäre, würde der Eigenname nicht segmentiert.)

Tokenisierung: Einzelfälle

- "1:1" wird auch als Abkürzung für die Wendung "eins zu eins" entsprechend der ausgeschriebenen Form segmentiert:

  ```
  1:1
  ```

  (Bei Fußballergebnissen u.ä. wird nach den Guidelines ebenfalls segmentiert.)

- Datumsangaben nach ISO 8601 werden analog zu "21/ 07/ 1980" semantisch in Jahr, Monat und Tag segmentiert. Also z.B. "1980-07-21" zu

  ```
  1980
  -07
  -21
  ```

- Kapitelnummern wie "2.1.3" oder "4.2." werden nicht segmentiert:

  ```
  2.1.3
  4.2
  ```

- Aufzählungen wie "1)", "(i)", "(1)" usw. werden in Nummer und Satzzeichen segmentiert (vgl. u.a. Tiger-Baumbank):

  ```
  1
  )
  ```

  bzw.

  ```
  (
  ```

```
i
)
```

Eine Ausnahme bilden lediglich Ordinalzahlen wie “1.”, die nach den STTS-Guidelines als ein Token behandelt werden:

```
1.
2.
```

- Im doppelt gemoppelten Sonderfall “a.)”, “b.)”, etc. folgen wir einer Mehrheitsentscheidung der Annotatoren und zerlegen jeweils in zwei Token:

```
a.
)
b.
)
```

- Analog zu genderneutralen Gruppenbezeichnern wie “Student(in)” werden optionale Pluralbildungen wie “Portemonnaie(s)” oder “Musikkultur(en)” nicht segmentiert:

```
Portemonnaie(s)
Musikkultur(en)
```

Dies entspricht dem Vorgehen in Tiger; das abgetrennte Pluralsuffix könnte auch nicht sinnvoll in STTS getaggt werden.

- Alle Doppeltoken mit eingeklammertem Präfix werden segmentiert, nicht nur solche mit Ergänzungsstrich; z.B. “(Nacht)zug” zu

```
(
Nacht
)
zug
```

- ISBN-Nummern werden analog zu einem komplexen Eigennamen interpretiert und bleiben damit als ein Token erhalten, z.B.

```
ISBN
3-89115-142-X
```

- Verschleierte E-Mail-Adressen werden wie normale E-Mail-Adressen als ein Token behandelt und entsprechend als EML getaggt:

```
schtepf[at]gmx[dot]net
```

Wie mit Fällen umzugehen ist, in denen die verschleierten Sonderzeichen durch Leerzeichen abgetrennt sind, muss noch festgelegt werden.

- Die Regeln zur Behandlung von versehentlich zusammengeschriebenen Wörtern in Abs. 4.4 der Tokenisierungsguidelines finden nur dann Anwendung, wenn der zweite Wortteil *nicht* durch Binnenmajuskel, Interpunktionszeichen o.ä. erkennbar abgegrenzt ist (vgl. dazu die Trennung von “winke@bochum” in Abs. 4.11 sowie “unsereKomm” in Abs. 4.6). So wird z.B. “Google+verbunden” zu

```
Google+
verbunden
```

sowie in Analogie zu der Abtrennung von Maßeinheiten “2008mitarbeiter” zu

```
2008
mitarbeiter
```

- Iterierte und zusammengesetzte Interpunktionszeichen werden immer dann als ein Token behandelt, wenn sie eine funktionale Einheit bilden (analog zu iterierten Ausrufe- und Fragezeichen sowie komplexen mathematischen Operatoren in Abs. 4.1 der Guidelines). Dies gilt für ASCII-Gedankenstriche “---” ebenso wie für Wiki-Links: “[[Startseite]]” wird segmentiert zu

  ```
  [[
  Startseite
  ]]
  ```

  Doppelte Klammern wurden im CMC-Subset hingegen als separate Interpunktionszeichen und nicht als iterierte Formen gedeutet, d.h. als zwei Klammernebenen und nicht als funktionale Einheit. “((Allg.Infos))” wurde so zu

  ```
  (
  (
  Allg.
  Infos
  )
  )
  ```

-