Medienlinguistische Methodik
Annotation

Arne Rubehn
Lehrstuhl für Multilinguale Computerlinguistik
Universität Passau

25.11.2025

Überblick

In dieser Sitzung widmen wir uns kurz einigen **Gesätzmäßigkeiten** in Korpora und dann der **Annotation** von wissenschaftlichen Daten**.**

Was besagt das Zipfsche Gesetz?

Was verstehen wir unter Annotationen?

Wie und wofür werden Daten annotiert?

Korpuslinguistik (cont.)

Setzen wir unsere Programmieraufgabe auf Kaggle fort.

Hierbei zeigen sich ein paar interessante (statistische) Eigenschaften von Korpora!

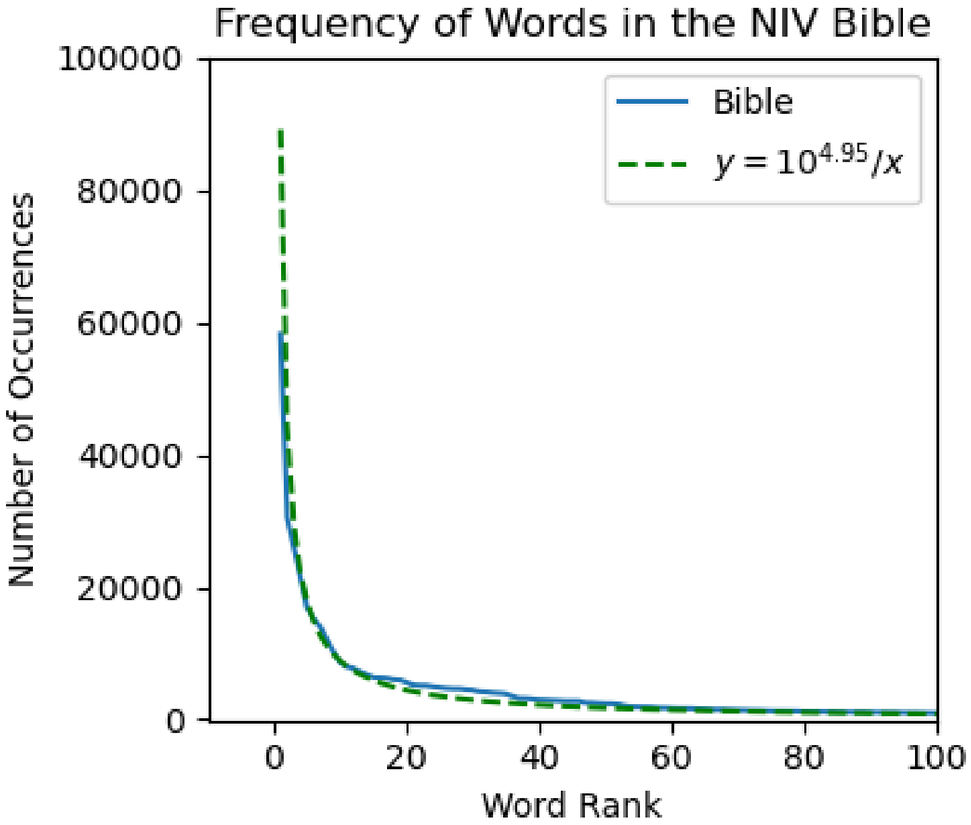
Zipfsches Gesetz

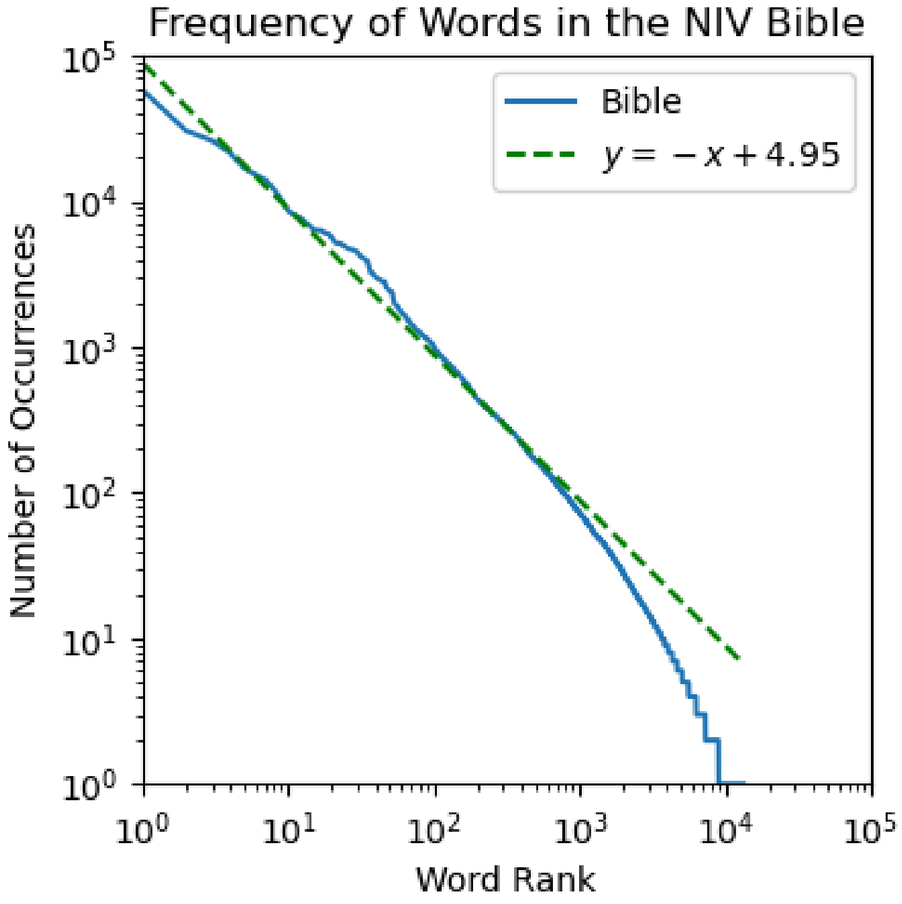
Das **Zipfsche Gesetz** besagt, dass die Frequenz eines Wortes in einem Korpus invers proportional zu seinem Rang (in einer "Rangliste" nach Frequenz) steht.

Weiß man also, auf welchem **Rang** ein Wort steht, kann man dessen **Frequenz** (in Abhängigkeit zu der des häufigsten Wortes) abschätzen – und umgekehrt auch. Sehr vereinfachte Faustformel: Das zweithäufigste Wort ist ½ mal so häufig wie das häufigste, das dritthäufigste ⅓ mal so häufig, etc...

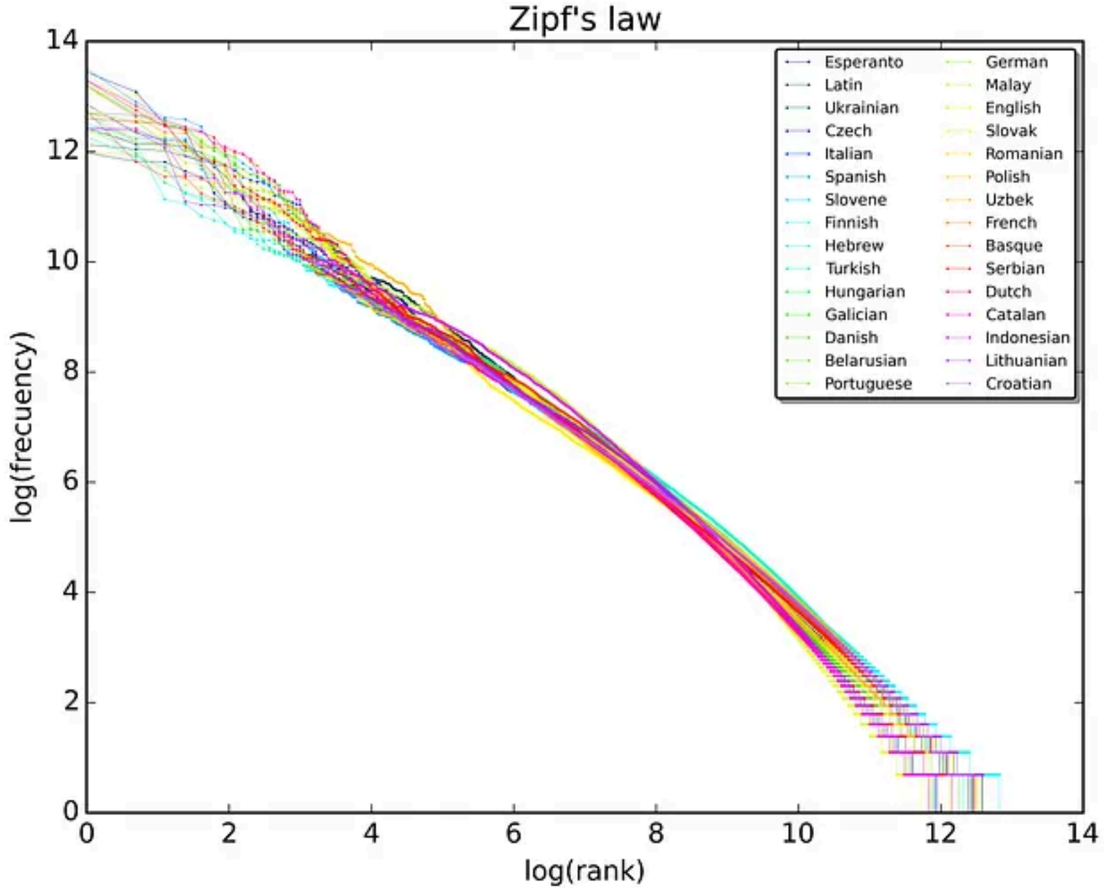
Für statistische Berechnungen werden Frequenzen und Ränge üblicherweise **logarithmisch** transformiert, da sich somit eine lineare Korrelation ergibt.

Zipfsches Gesetz





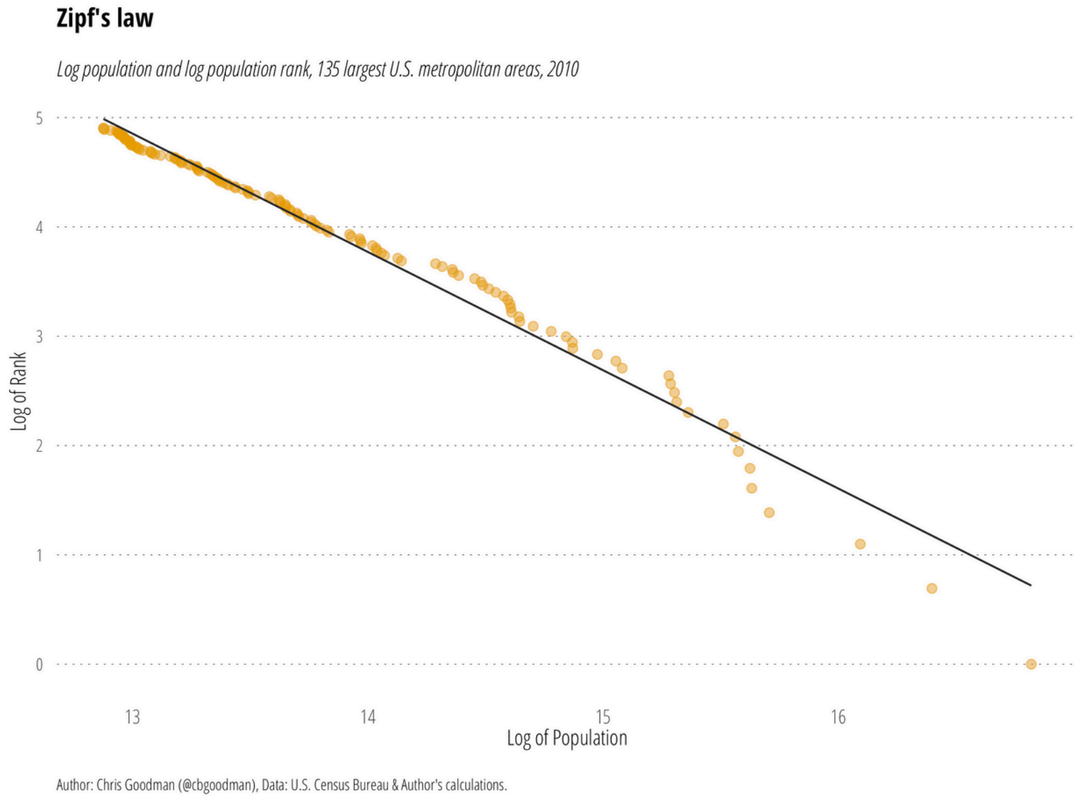
Zipfsches Gesetz



Zipfsches Gesetz

Das Zipfsche Gesetz gilt auch für viele Verteilungen außerhalb der Linguistik, z.B. die Größe von Städten, Tierkommunikation oder Unternehmensbeziehungen.

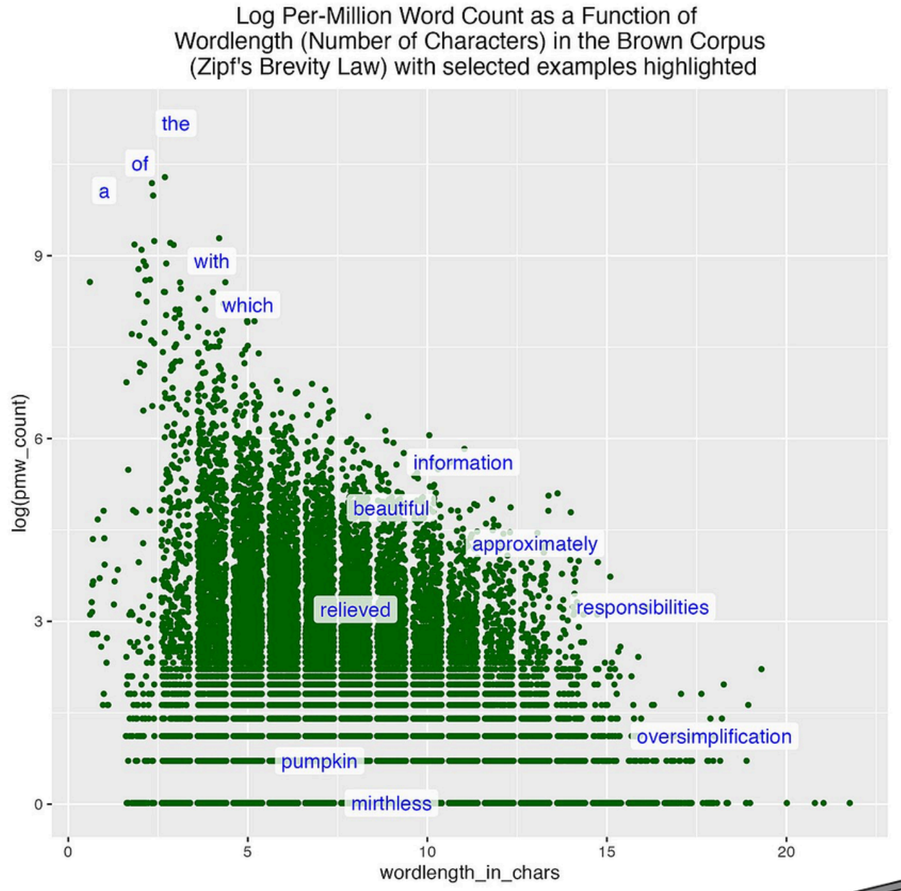
Verwandte Effekte sind z.B. das **Pareto-Prinzip** (auch bekannt als 80-20-Regel).



Zipfsches Gesetz der Abkürzung

Eine verwandte – aber nicht identische – Gesetzmäßigkeit ist das **Zipfsche Gesetz der Abkürzung** *(Zipf's Law of Abbreviation)*.

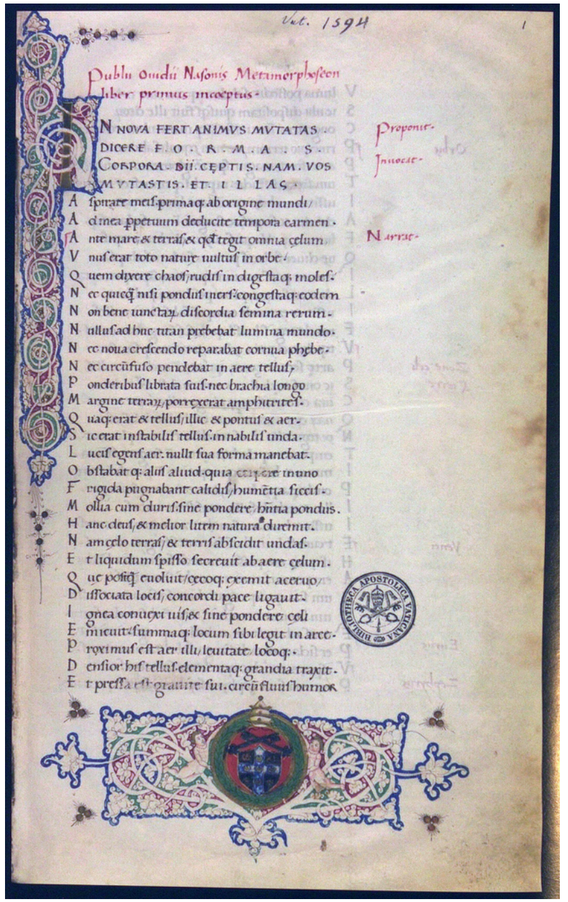
Demnach sind hochfrequente Wörter tendenziell kürzer als seltene Wörter. Dieser Effekt ist über sämtliche Sprachen hinweg stabil und trifft auch z.B. auf Ortsnamen zu.



Annotation

Unter **Annotation** versteht man ganz allgemein das **Anreichern** oder **Strukturieren von Informationen** bzw. Daten.

Im weitesten Sinne kann jede Form der “Zusatzinformation” als Annotation verstanden werden: Kommentare in historisch-kritischen Ausgaben, handgeschriebene Notizen am Rand von Texten, Beschriftungen von Fotos, …



Annotation

Unter **Annotation** versteht man ganz allgemein das **Anreichern** oder **Strukturieren von Informationen** bzw. Daten.

Im weitesten Sinne kann jede Form der “Zusatzinformation” als Annotation verstanden werden: Kommentare in historisch-kritischen Ausgaben, handgeschriebene Notizen am Rand von Texten, Beschriftungen von Fotos, …

Annotation

Im etwas engeren Sinne verstehen wir unter Annotation das gezielte Erstellen von **strukturierten Daten.** Dem liegt natürlich ein spezifisches Datenmodell zugrunde.

Hierbei geht es darum, implizite Informationen **explizit** und **maschinenlesbar** darzustellen. In diesem Sinne wird meistens die eigentliche **Annotation** (das Anreichern von Daten) von **Metadaten** (Daten *über* die Daten) unterschieden.

“Measure what is measurable, and make measurable what is not so.”

Annotation: CoNLL-U

```
# sent_id = 1
# text = They buy and sell books.
1   They     they     PRON    PRP    Case=Nom|Number=Plur                2   nsubj   2:nsubj|4:nsubj   _
2   buy      buy      VERB    VBP    Number=Plur|Person=3|Tense=Pres     0   root    0:root            _
3   and      and      CCONJ   CC     _                                   4   cc      4:cc              _
4   sell     sell     VERB    VBP    Number=Plur|Person=3|Tense=Pres     2   conj    0:root|2:conj     _
5   books    book     NOUN    NNS    Number=Plur                         2   obj     2:obj|4:obj       SpaceAfter=No
6   .        .        PUNCT   .      _                                   2   punct   2:punct           _
```

Annotation: CoNLL-U

Metadaten

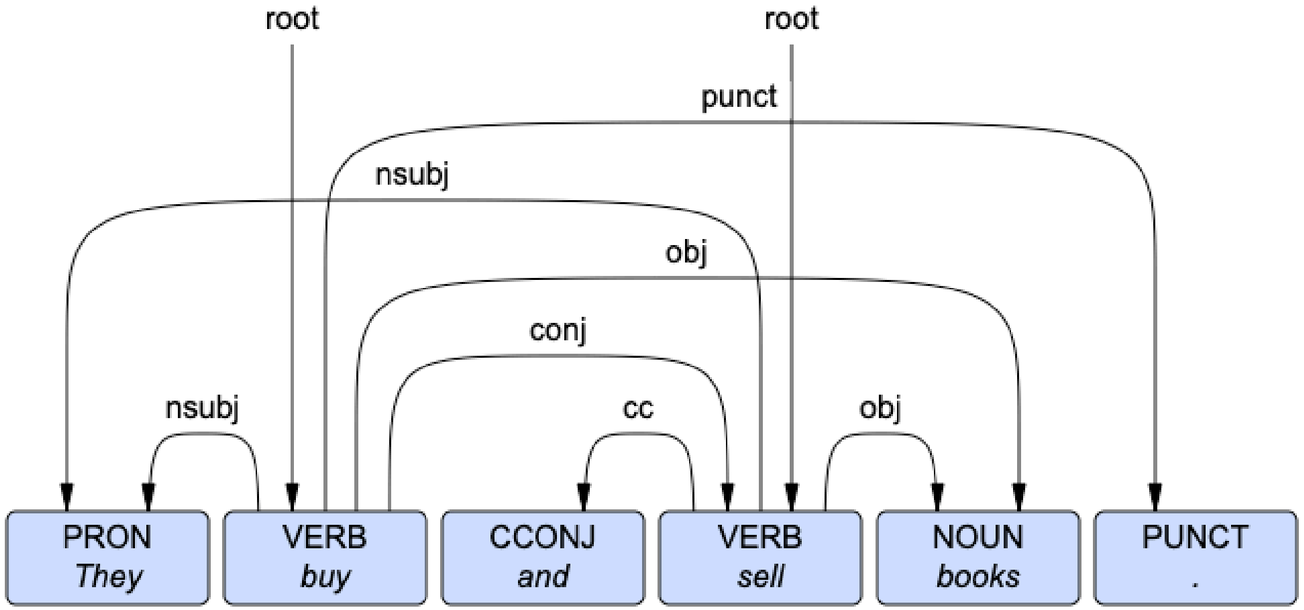
Primärdaten

```
# sent_id = 1
# text = They buy and sell books.
1   They     they     PRON    PRP   Case=Nom|Number=Plur                  2   nsubj   2:nsubj|4:nsubj   _
2   buy      buy      VERB    VBP   Number=Plur|Person=3|Tense=Pres       0   root    0:root            _
3   and      and      CCONJ   CC    _                                     4   cc      4:cc              _
4   sell     sell     VERB    VBP   Number=Plur|Person=3|Tense=Pres       2   conj    0:root|2:conj     _
5   books    book     NOUN    NNS   Number=Plur                           2   obj     2:obj|4:obj       SpaceAfter=No
6   .        .        PUNCT   .     _                                     2   punct   2:punct           _
```

Annotation

Annotation: CoNLL-U

```
# sent_id = 1
# text = They buy and sell books.
1   They     they     PRON    PRP    Case=Nom|Number=Plur                2   nsubj   2:nsubj|4:nsubj   _
2   buy      buy      VERB    VBP    Number=Plur|Person=3|Tense=Pres     0   root    0:root            _
3   and      and      CCONJ   CC     _                                   4   cc      4:cc              _
4   sell     sell     VERB    VBP    Number=Plur|Person=3|Tense=Pres     2   conj    0:root|2:conj     _
5   books    book     NOUN    NNS    Number=Plur                         2   obj     2:obj|4:obj       SpaceAfter=No
6   .        .        PUNCT   .      _                                   2   punct   2:punct           _
```



Annotation: CoNLL-U

Metadaten

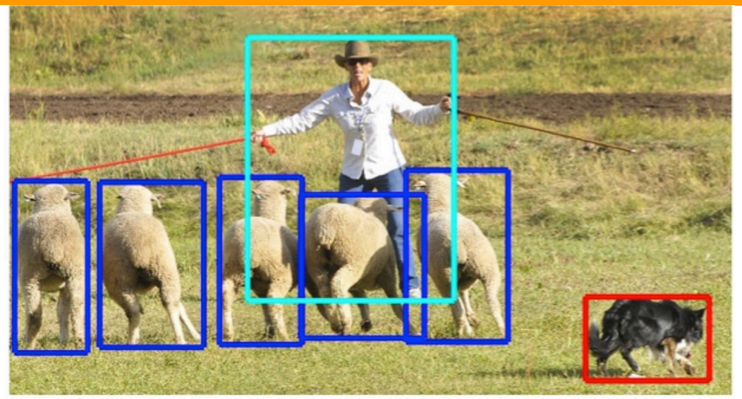
Primärdaten

```
# sent_id = panc0.s4
# text = तत् यथानुश्रूयते।
# translit = tat yathānuśrūyate.
# text_fr = Voilà ce qui nous est parvenu par la tradition orale.
# text_en = This is what is heard.
1       तत्             तद्     DET     _       Case=Nom|…|PronType=Dem     3       nsubj   _       Translit=tat|LTranslit=tad|Gloss=it
2-3     यथानुश्रूयते   _       _       _       _                           _       _       _       SpaceAfter=No
2       यथा             यथा    ADV     _       PronType=Rel                3       advmod  _       Translit=yathā|LTranslit=yathā|Gloss=how
3       अनुश्रूयते     अनु-श्रु VERB    _       Mood=Ind|…|Voice=Pass       0       root    _       Translit=anuśrūyate|LTranslit=anu-śru|Gloss=
4       ।               ।       PUNCT   _       _                           3       punct   _       Translit=.|LTranslit=.|Gloss=.
```

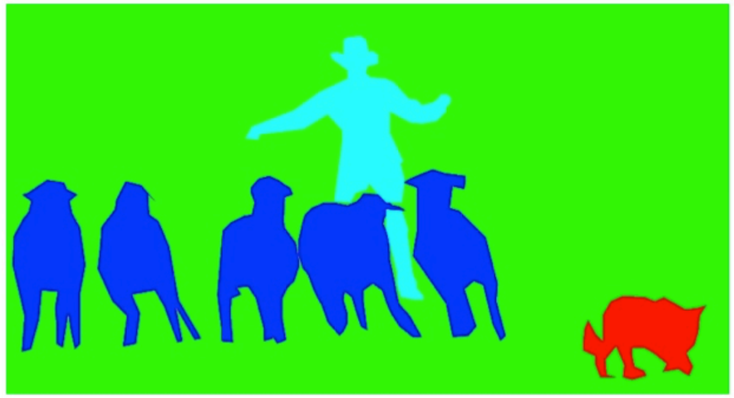
Annotation

Annotation: COCO

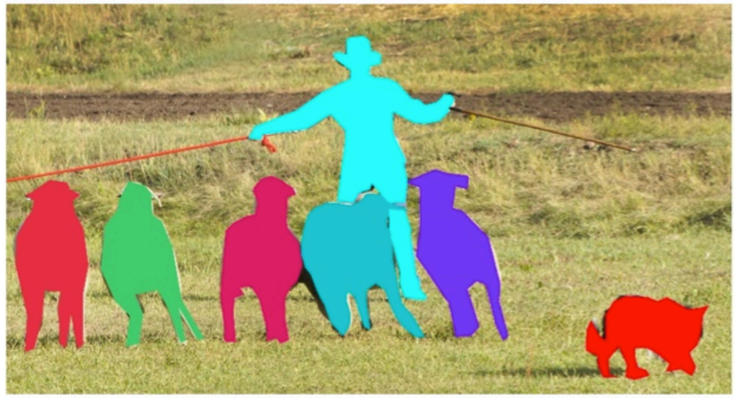
(a) Image classification



(b) Object localization

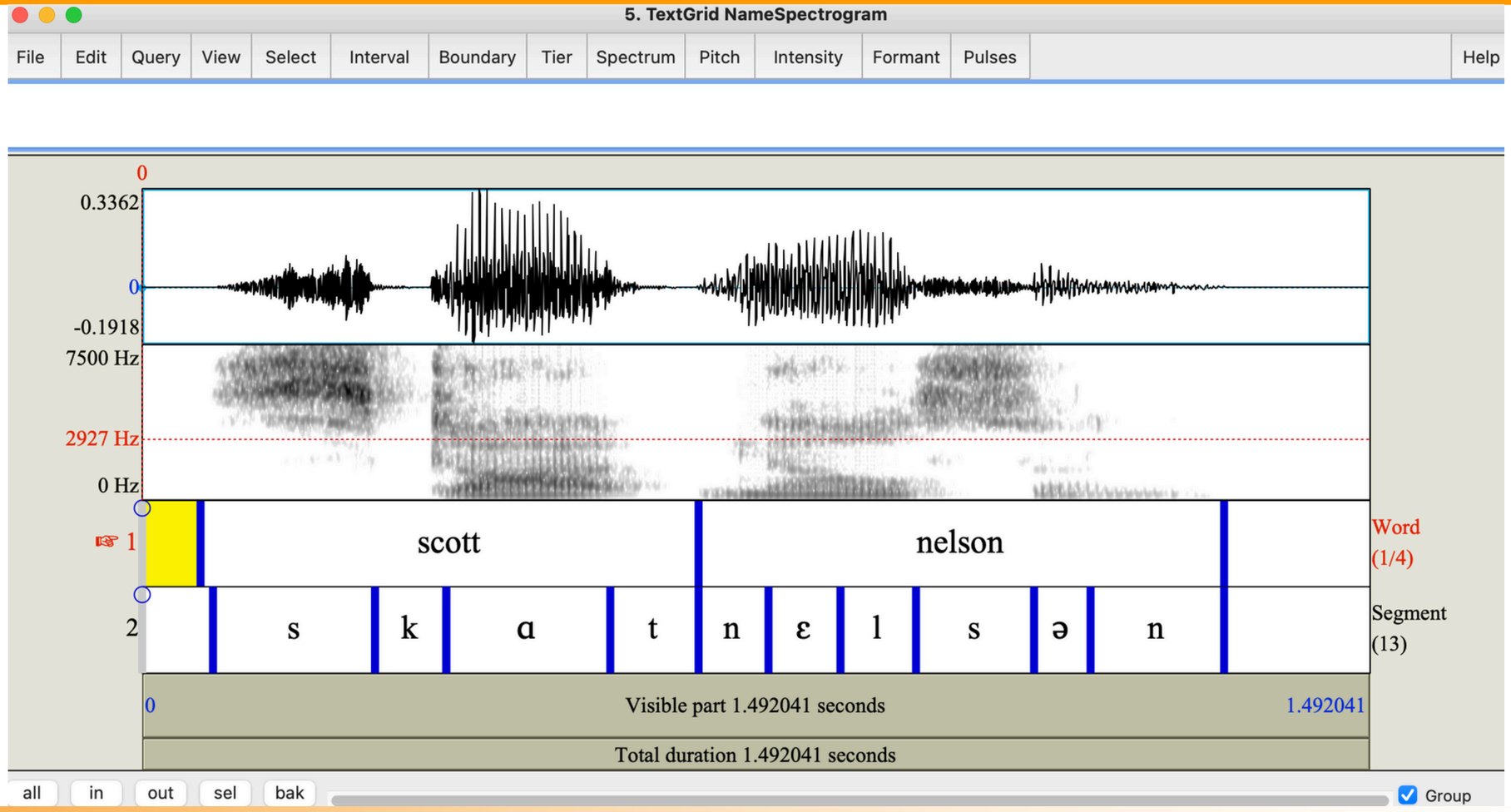


(c) Semantic segmentation



(d) This work

Annotation: Praat



Annotation

Wie wir sehen, können Daten anhand von Annotationen um unterschiedlichste Zusatzinformationen erweitert werden.

Je nach **Modalität** der Primärdaten (Text, Audio, Bild, …) und **Ziel** der Annotation sind verschiedene Repräsentationen notwendig.

Sehr häufig werden bei Annotationen Teile der Primärdaten **kategorisiert,** also einem vordefinierten Inventar an Tags zugeordnet (daher auch **Tagging** genannt).

Annotationen können auf verschiedenen **Ebenen** stattfinden.

Annotation

Die Annotation von Daten kann grob gesagt zu zwei Zwecken erfolgen.

Einerseits können **Zusatzinformationen für Leser:innen** bereitgestellt werden, die das Verständnis der Daten erleichtern.

Andererseits werden für das **Training von Computermodellen** häufig größere Mengen an hochstrukturierten & annotierten Daten benötigt.

In beiden Fällen gilt:

Eine korrekt ausgeführte Annotation ändert die Primärdaten nicht!

Annotation: In eigener Sache

Auf https://tvl.digling.org findet ihr eine digitale Edition von Sprachdaten aus Vanuatu, an der ich gerade arbeite.

Untersucht die dargestellten Daten:

1. Wie werden Primärdaten referenziert?
2. Welche Annotationsschritte wurden durchgeführt?
3. Welche Metainformationen wurden beigefügt?

Annotation: NER-Tagging

Eine klassische Aufgabe für Computermodelle in der Sprachverarbeitung ist die **Named Entity Recognition (NER),** also die Erkennung von Eigennamen in Texten.

Für das Training solcher Modelle sind annotierte Trainingsdaten erforderlich.

Annotiere die auf StudIP hochgeladene Textdatei mit dem webbasierten NER-Annotationstool unter https://arunmozhi.in/ner-annotator/. Exportiere deine Annotationen als JSON und lade sie auf StudIP unter "Studienleistungen" hoch.

Überlege dir vorher, welche Kategorien (Name, Ort, ...) du annotieren solltest!

Annotation: Evaluation

Um gute Annotationen zu erzeugen, werden wissenschaftliche Daten typischerweise von **mehreren Annotator:innen** bearbeitet.

Um eine gute Qualität zu gewährleisten, sollten die Annotationen der unterschiedlichen Personen möglichst gut **übereinstimmen.** Je nach Art der Annotation und Anzahl der Annotator:innen gibt es verschiedene statistische Maße, um das ***Inter-Annotator-Agreement (IAA)*** zu bestimmen.

Annotation erfolgt **zyklisch**: Guidelines werden basierend auf IAA (und spezifischen Problemen) angepasst; Daten werden anhand der neuen Guidelines erneut annotiert.