Evaluation of Annotation Strategies

Bachelorarbeit

Natural Language Processing (NLP)

- Ziel: Kommunikation zwischen Mensch und Maschine
- Analyse von geschriebener/gesprochener Sprache
- Probleme: Mehrdeutigkeit und Komplexität der Sprache
- Lösung: Annotation

Annotation und Internet

- Zusätzliche Informationen zum Text
- Riesige Anzahl von Daten im Internet
- Computer kann Sprache nicht verstehen
- Metadaten müssen hinzugefügt werden → Annotation

Arten von Annotation

- Part of Speech Tagging
- Eigennamenerkennung
- etc.

Part of Speech (POS) Tagging

- Part of Speech (Nomen, Adjektiv, etc.) wird Termen zugeordnet
- Verwendung: Suche von grammatikalischen/lexikalischen Mustern ohne Wörter zu definieren
- Zum Beispiel: Finde alle Nomen im Plural ohne vorangehenden Artikel

- Sammlung von POS-Tags → Tagset
- Stuttgart Tübigen Tagset ermöglicht Wiederverwendung von annotierten Texten
- 54 Tags mit 11 Hauptwortarten

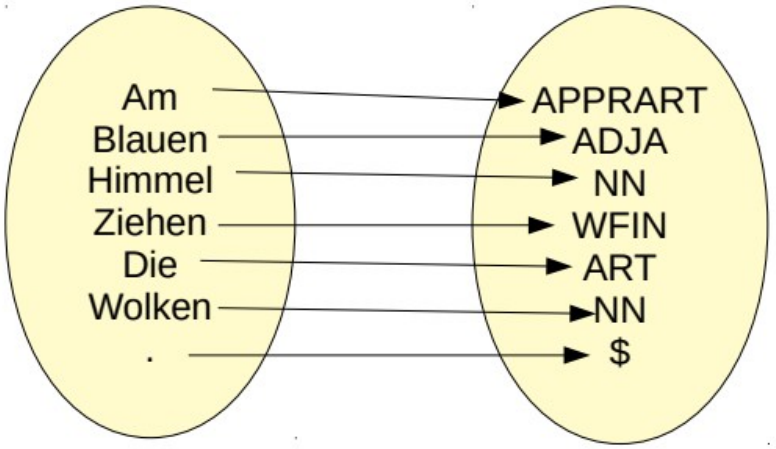


Abbildung 1: Beispiel für Pos-Tagging [1]

- Schwierigkeiten:
  - Out of Vocabulary Fälle: Unbekannte Wörter
  - Mehrdeutigkeit: Wort kann mehrere Tags haben, Sucht → Verb (VVFIN) vs. Nomen (NN)
- Einer der meist verwendeten Methoden
- Von Computer-Programmen ausführbar
  - Taggit: 71% Genauigkeit
  - CLAWS: 91% Genauigkeit

Eigennamenerkennung

- Durch MUC und CoNLL Konferenzen im Zentrum
- Klassifikation von Eigenname durch MUC:
  - Personen
  - Unternehmen
  - Geographische Ausdrücke
  - Datums- und Maßangaben
- Erzielt im Englischen 95% Korrektheit im Deutschen schlechter

- Probleme im Deutschen: Eigennamen nicht durch Groß- und Kleinschreibung erkennbar
- Regeln nicht anwendbar
- z.B. ein Vorname gefolgt von einem großgeschriebenem Wort ist ein Personenname (schreibt Lisa Bücher)

irrtümlich als Eigenname erkannt

Manuelle vs. Semi-automatische Annotation

- Manuell: Person annotiert Textkorpus
  - Sprachkenntnis beeinflusst Ergebnis stark
- Semi-automatisch: Person korrigiert automatisch generierte Annotation
  - Zeit Ersparnis
  - Qualitativ hochwertigeres Ergebnis

Semi-automatisch Annotation

- Datensatz muss erstellt werden
  - Mittels Distant/Weak Supervision automatisch generierbar
- Probleme bei Distant/Weak Supervision
  - Labels werden falsch vergeben
  - Führt zu Performance Verlust

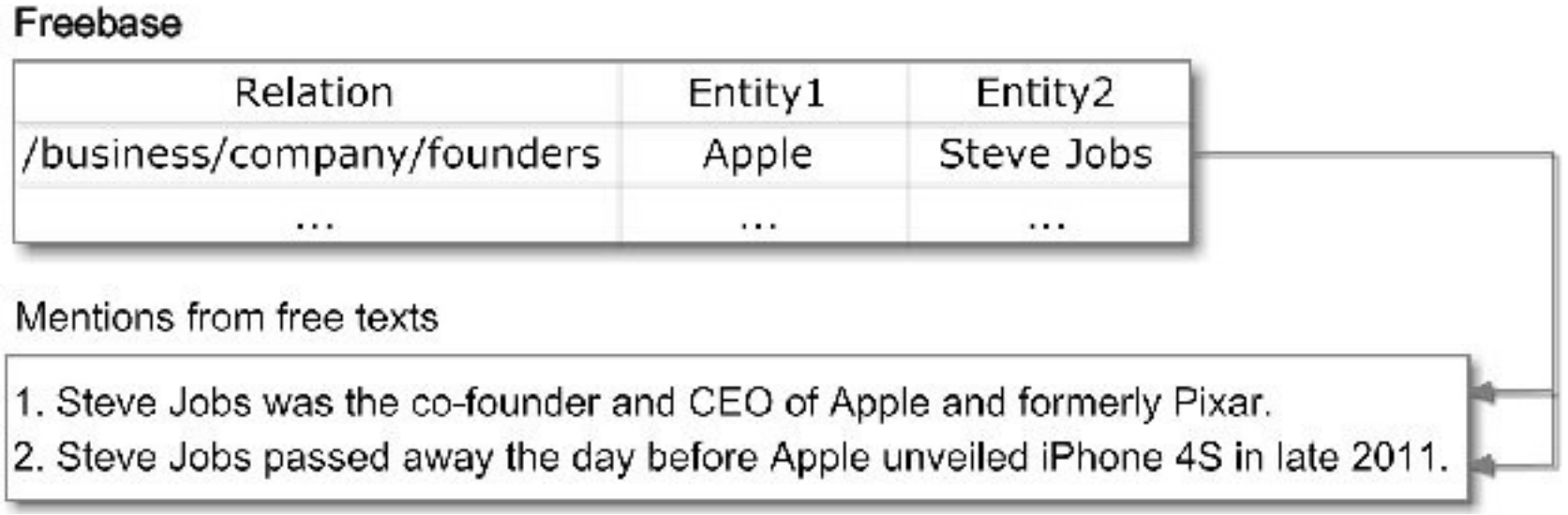


Abbildung 2: fehlerhafter Datensatz erstellt mit Distant Supervision [2]

Annotationstools

- Zahlreiche verfügbare Tools
- Unterschied in:
  - Funktion
  - Aussehen
  - viele weitere Aspekte

Brat

- Online nutzbar
- Selbsterklärenden Oberfläche
- Verschiedene Sprachen und Annotations-Typen werden unterstützt
- Vertraute Umgebung
- Mittels Mausklick Label hinzufügen, Beziehungen mittels drag&drop
- Annotations-Zeit um 15% verringert

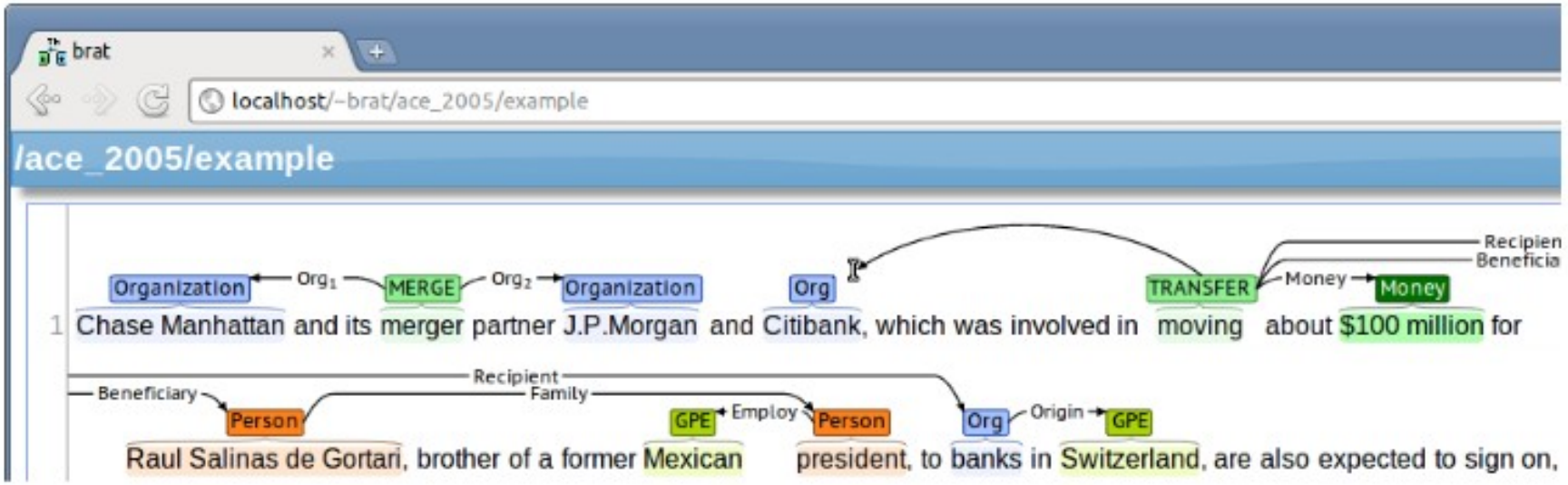


Abbildung 3: Annotation mittels Brat [3]

BioQRator

- Online nutzbar
- Für medizinische Literatur
- Erste Web-Tool mit Unterstützung von BioC Format
  - Einfaches Format für Austausch von Dokumenten
  - In Absätzen mit zusätzlichen Annotationen
- Intuitive Benutzeroberfläche per Mausklick Labels und Beziehungen

Evaluation

CodeAnnotator

- Webtool
- Entwickelt am Know-Center TUGraz
- Datensätze können erstellt werden und Dokumente hinzugefügt werden
- Zwei Methoden zur Annotation:
  - Manuelle Annotation: Annotation mit Brat Editor
  - Keyword in Context: Suchbegriff wir im Kontext angezeigt und kann akzeptiert oder verworfen werden

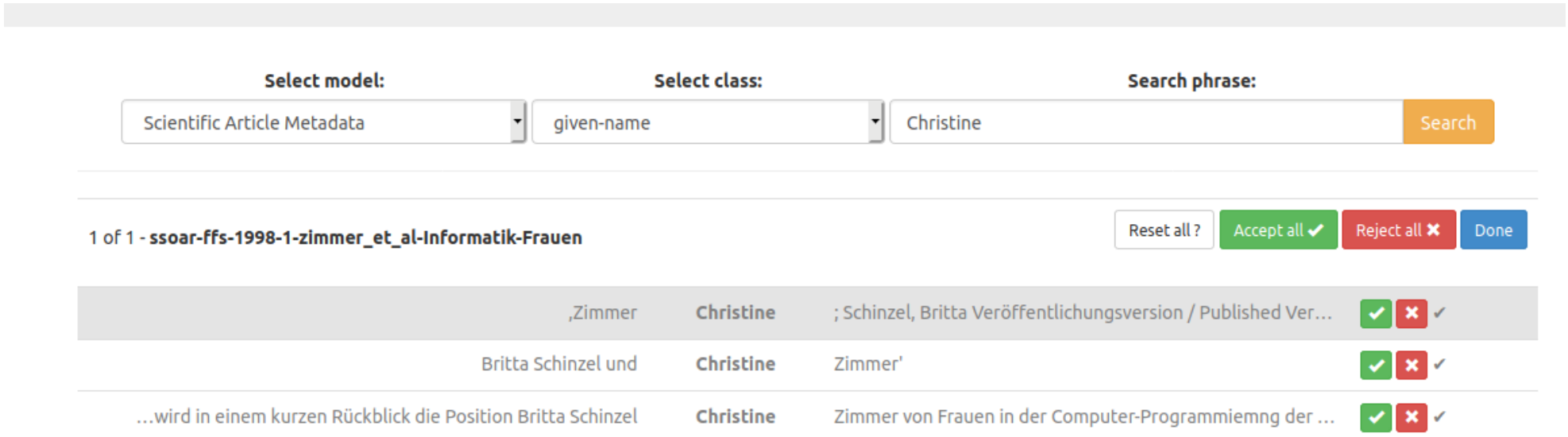


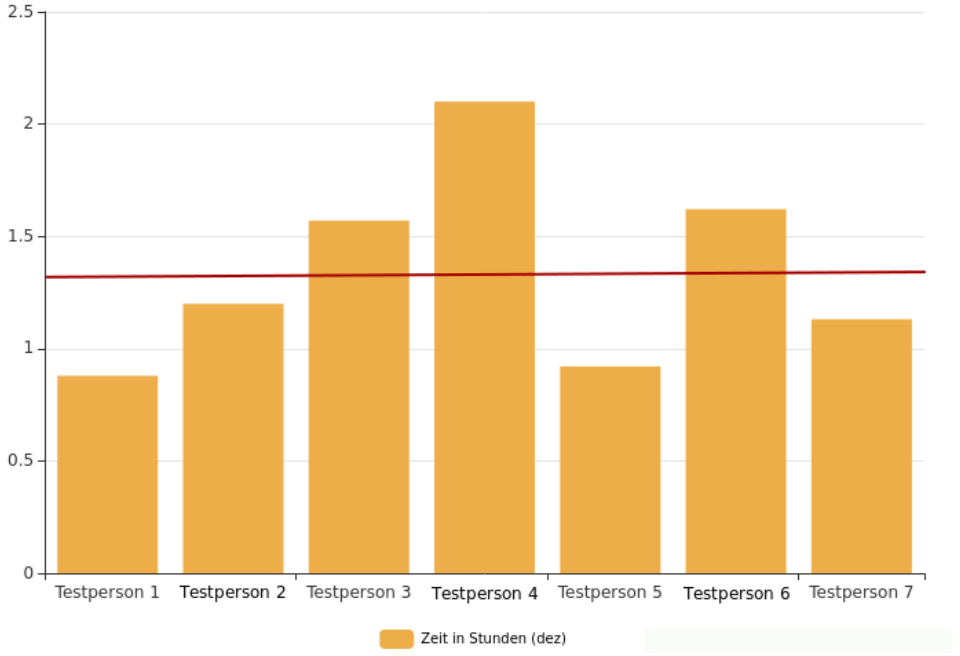
Abbildung 4: Keyword in Context Annotation [4]

Evaluation Vorbereitung und Ablauf

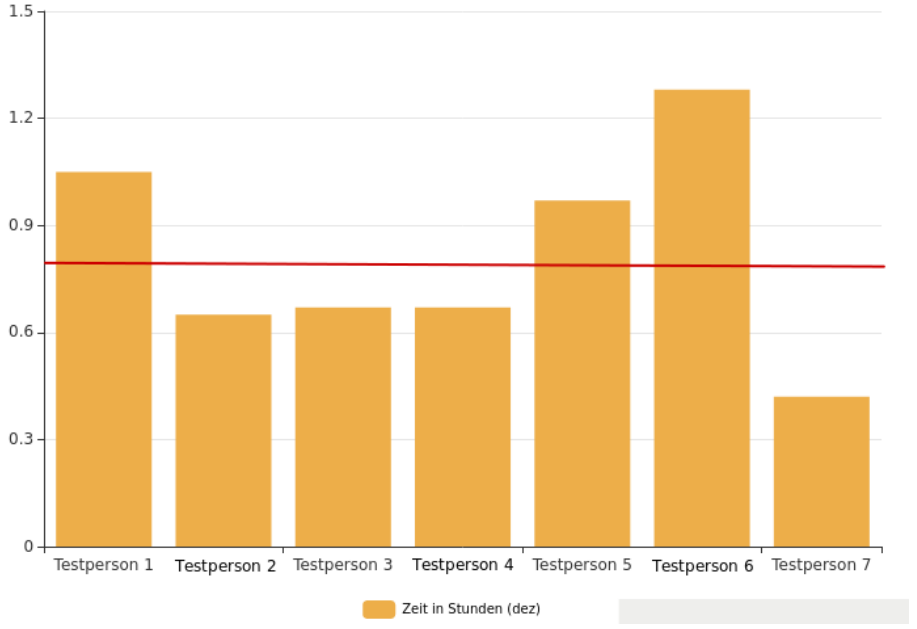
- 7 Teilnehmer annotieren jeweils mit beiden Methoden
- 2 wissenschaftliche Dokumente zur Annotation
  - Dokument 1: Informatik-Frauen
  - Dokument 2: Informatik – Kompetenzentwicklung bei Kindern

- Modell für wissenschaftliche Dokumenten:
  - title
  - journal
  - subtitle
  - academic-title
  - given-name
  - middle-name
  - surname
  - index
  - affiliation
  - email
  - doi
  - ref-authorGivenName
  - ref-authorSurname
  - ref-authorOther
  - ref-editor
  - ref-title
  - ref-date
  - ref-publisher
  - ref-issueTitle
  - ref-bookTitle
  - ref-pages
  - ref-location
  - ref-conference
  - ref-source
  - ref-volume
  - ref-edition
  - ref-issue
  - ref-url
  - ref-note
  - ref-other

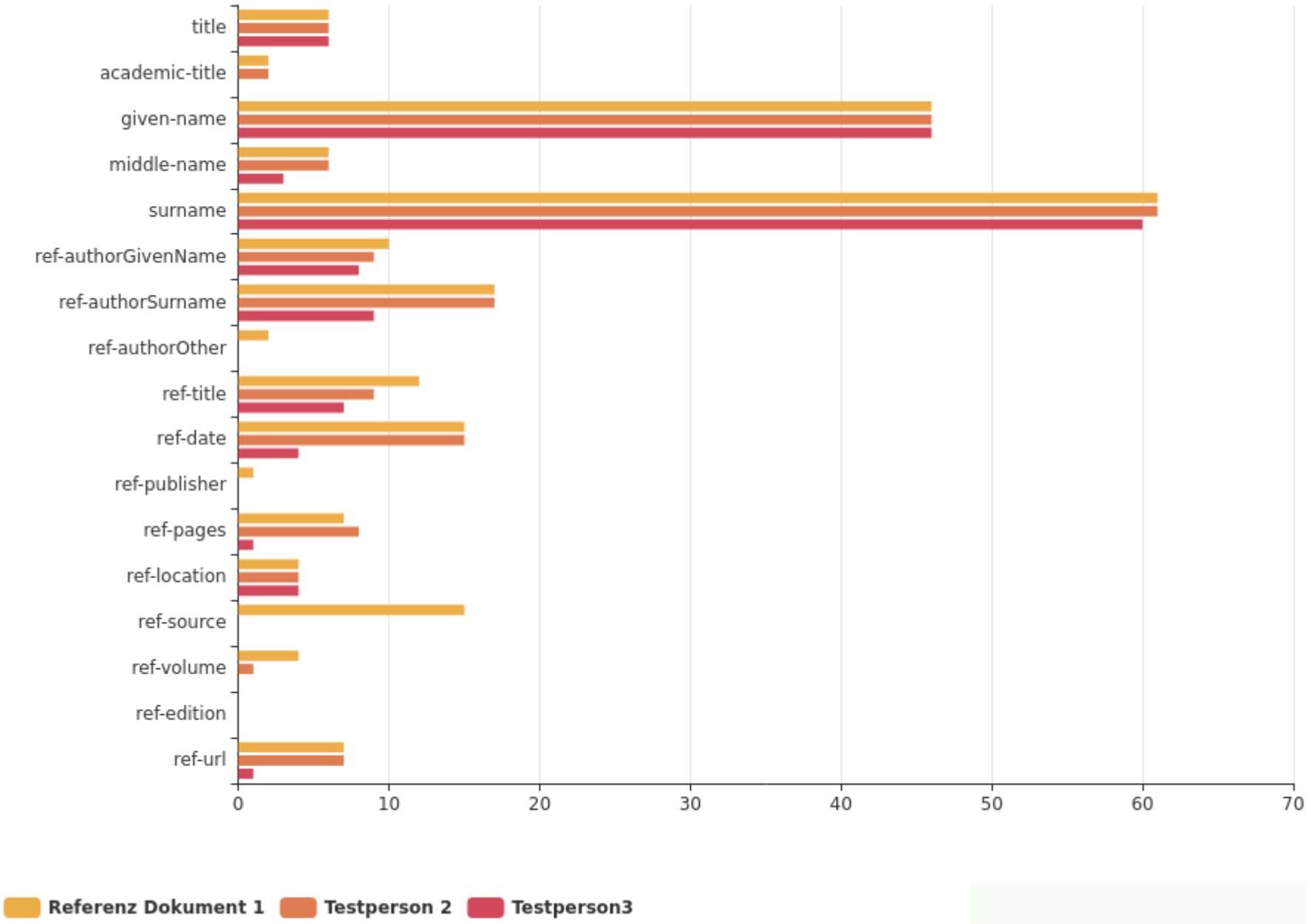
Ergebnis und Diskussion



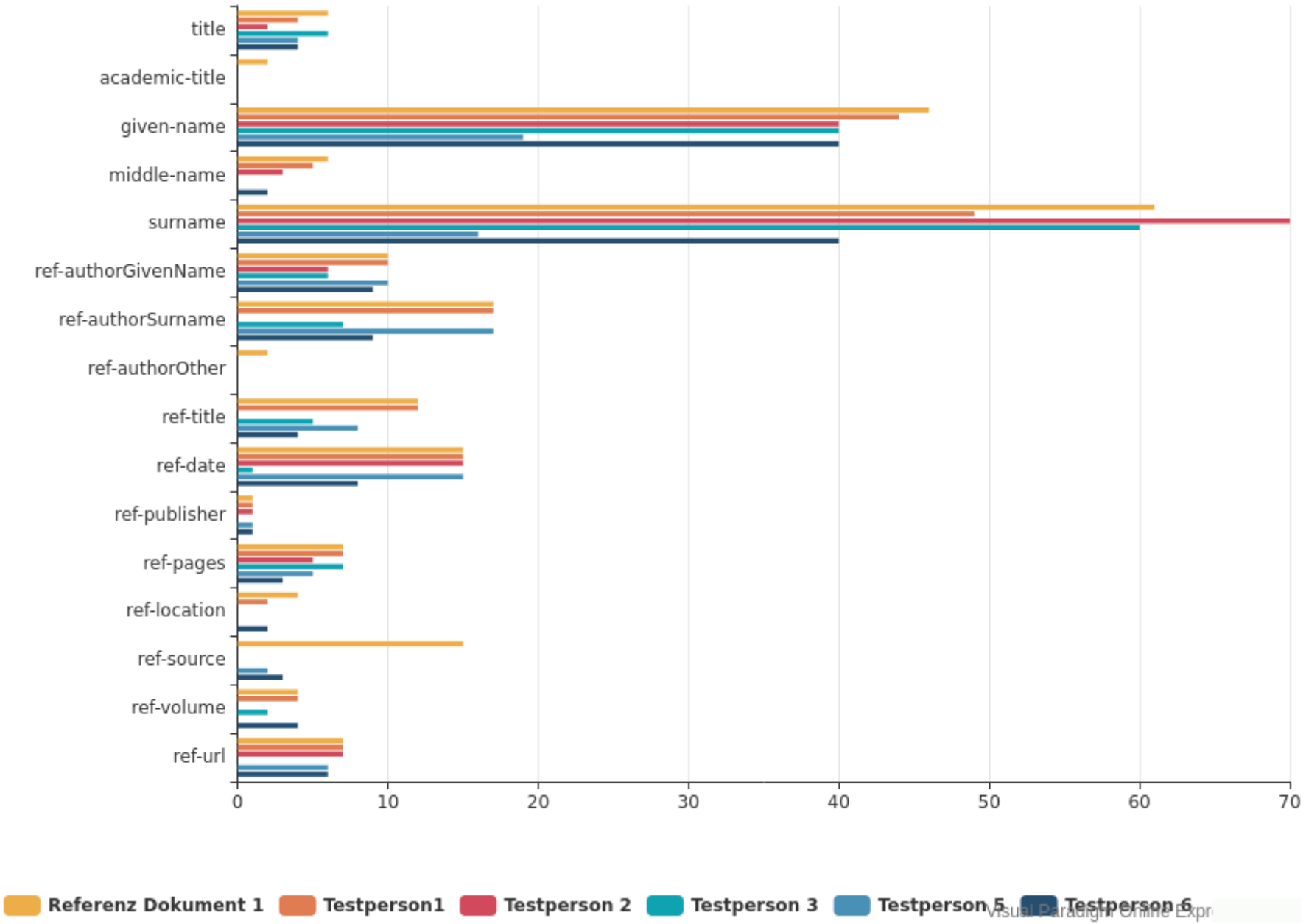
Benötigte Zeit bei Methode 1
inklusive Durchschnitt (dezimal)



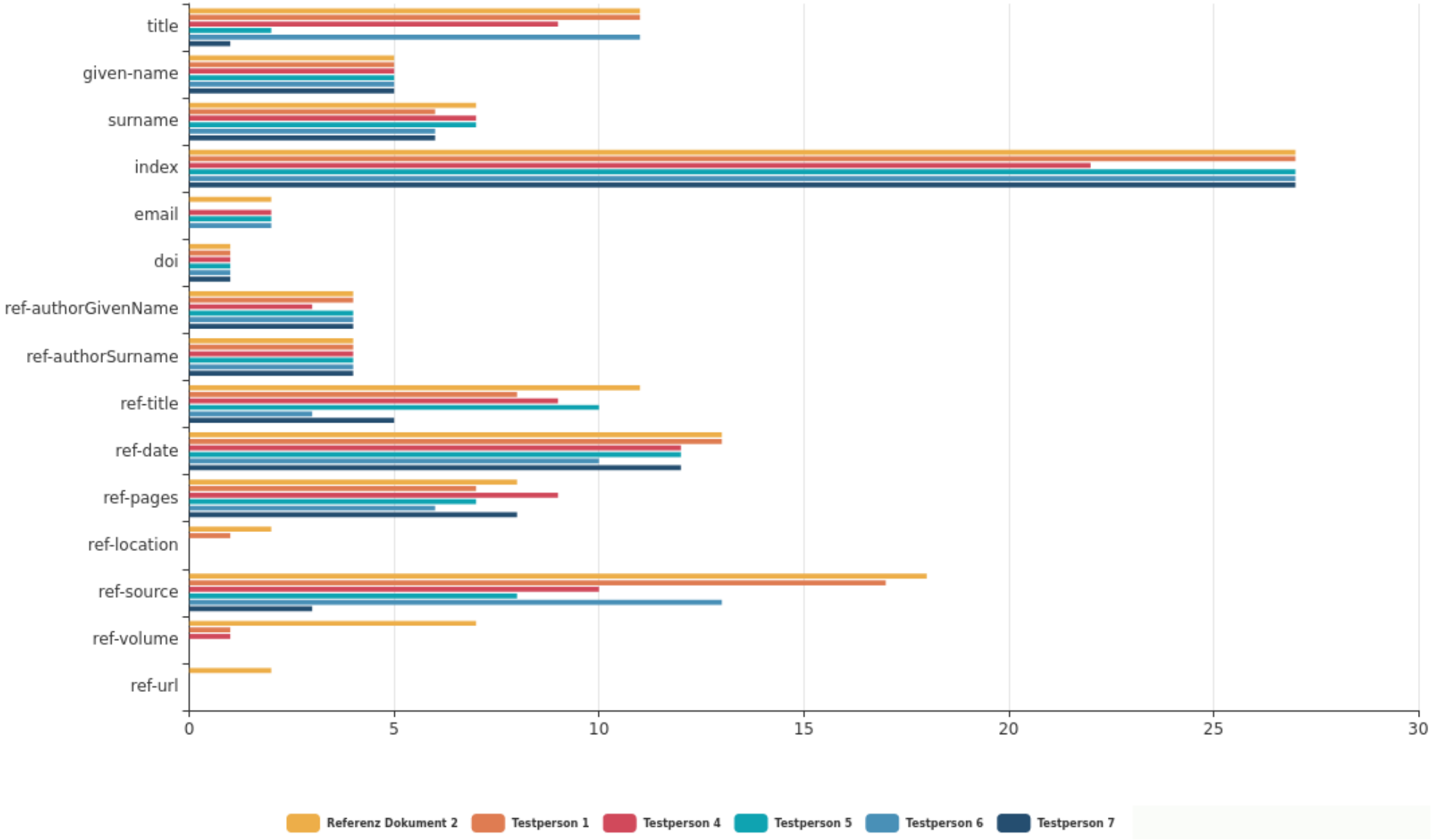
Benötigte Zeit bei Methode 2
inklusive Durchschnitt (dezimal)



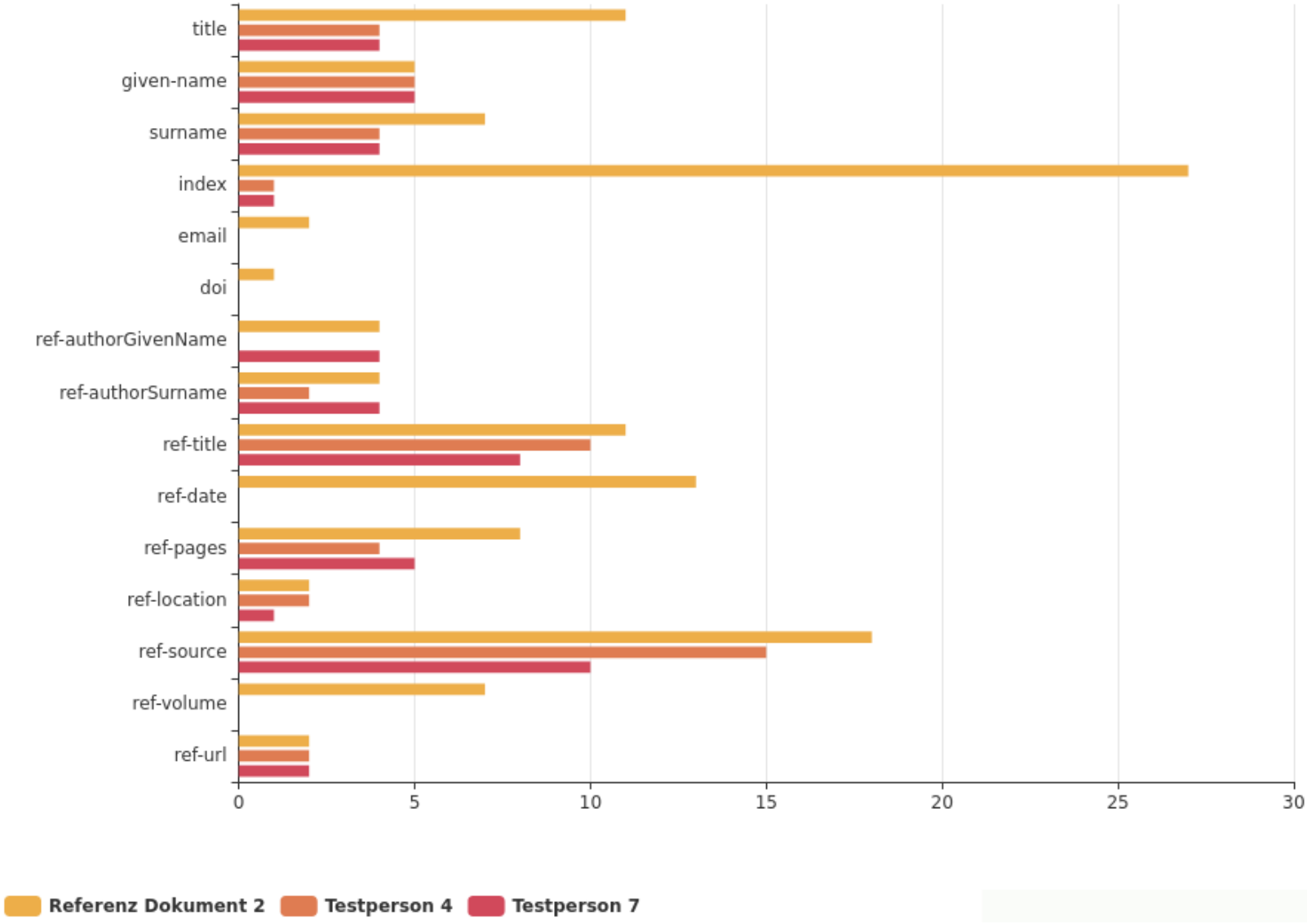
Anzahl von richtig Annotierten Wörtern der Einzelnen Klassen bei Dokument 1 mit Methode 1



Anzahl von richtig Annotierten Wörtern der Einzelnen Klassen bei Dokument 1 mit Methode 2



Anzahl von richtig Annotierten Wörtern der Einzelnen Klassen bei Dokument 2 mit Methode 1



Anzahl von richtig Annotierten Wörtern der Einzelnen Klassen bei Dokument 2 mit Methode 2

- Brat genauer, Dauer dafür länger
- Eindeutige Klassen (Vorname, Nachname) häufig richtig

Selbsteinschätzung

- Methode 1: Selbsteinschätzung schlechter als tatsächliches Ergebnis
- Methode 2: Selbsteinschätzung besser als tatsächliches Ergebnis
- Grund: fehlende Erfahrung, kein direktes Feedback
- 4 von 7 würden Tool nicht privat nutzen (kein Bedarf)
- 6 von 7 Teilnehmer finden KWIC Methode besser

Quellenangabe

- [1] U. Reichel. Sprachsynthese: Part-of-speech-tagging, 2016.
- [2] D. Zeng, K. Liu, Y. Chen, and J. Zhao. Distant supervision for relation ex-traction via piecewise convolutional neural networks. In Proceedings of the 2015 conference on empirical methods in natural language processing, pages 1753–1762, 2015.
- [3] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. brat:a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 102–107, Avignon, France, Apr. 2012. Association for Computational Linguistics.
- [4] i. p. d. w. t. C. p. Know-Center GmbH. Code annotator tool, 2012-2014.