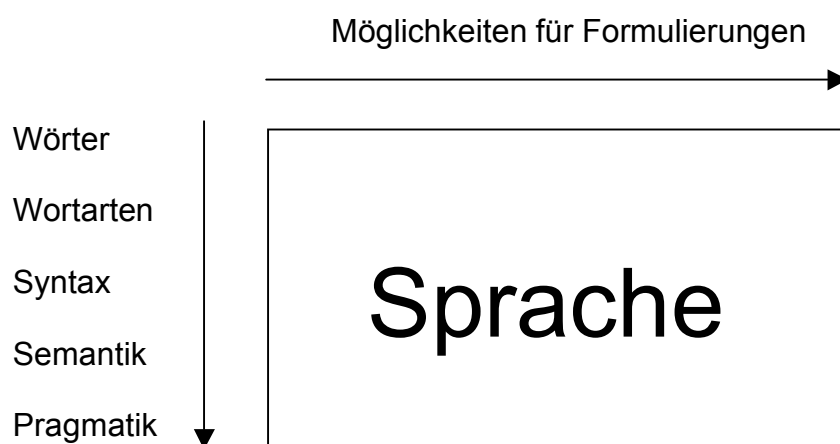


Übungsaufgaben 3

- 3.1. Betrachten Sie die folgende Grafik, die den “Sprachraum” zweidimensional nach “Tiefe” (der linguistischen Ebene) und “Breite” (die Menge aller möglichen Phänomene) aufspannt.



Zeichnen Sie ein, welchen Teil dieses “Sprachraumes” nach Ihrem Verständnis (a), introspektive “Armchair”-Modelle und (b), empirische Korpusmodelle typischerweise abdecken. Begründen Sie jeweils in einem oder zwei Sätzen.

- 3.2. Warum ist es problematisch, die manuelle Annotation von Korpora durch automatische Methoden (z.B. Vorschläge für den Annotator) unterstützen zu lassen? Nennen Sie mindestens zwei Schwierigkeiten mit kurzer Erklärung.
- 3.3. Beim Aufstellen eines Annotationsschemas muss auch entschieden werden, wie *detailliert* die zu annotierenden Kategorien sein sollen. Beschreiben Sie jeweils in wenigen Worten die Auswirkungen dieser Entscheidung auf Annotation *und* Nutzung des Korpus.
- 3.4. Wie Sie in der Vorlesung gehört haben, sind alle verfügbaren Korpora mit Annotation auf „tiefen“ linguistischen Ebenen deutlich kleiner als entsprechende Korpora mit „flacher“ Annotation. Weshalb ist das vermutlich der Fall? Was heißt das für die empirische Modellierung tieferer linguistischer Ebenen?
- 3.5. Nehmen Sie an, Sie hätten einen Parser (d.h. ein Modell für die syntaktische Struktur von Sätzen), der auf einem Zeitungskorpus trainiert sei. Die Korrektheit des Parsers wurde auf einem anderen Zeitungskorpus mit 90% bestimmt. Sie verwenden ihn, um Transkriptionen von Telefongesprächen (z.B. das Switchboard-Korpus) syntaktisch zu analysieren.
- 3.5.1. Spekulieren Sie kurz, wieso der Parser vermutlich schlechter funktioniert.
- 3.5.2. Nennen Sie zwei konkrete Phänomene in Switchboard, die dem Parser Probleme bereiten könnten. (Hinweis: Das Switchboard User’s Manual befindet sich unter http://www ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html)