

8. Lexik

Die Lexik ist die Disziplin, die sich mit der Wortebene beschäftigt. Unter dem **Lexikon** einer Sprache versteht man den Wortbestand der gesamten Sprache, während man den Wortbestand eines Texts oder den Wortschatz eines Sprechers/Hörers als **Vokabular** bezeichnet. Allerdings wird diese Unterscheidung nicht immer eingehalten. Der Terminus Lexikon steht in der Linguistik auch allgemein für ein Verzeichnis von bedeutungstragenden Einheiten (z.B. Morphem-Lexikon, Eigennamen-Lexikon).

Mit dem Lexikon beschäftigen sich zwei Disziplinen: die Lexikographie und die Lexikologie.

- Die **Lexikographie** hat die Aufgabe, den Wortbestand einer Sprache zu beschreiben. Dies geschieht in der Regel mittels Wörterbüchern¹. Es gibt eine Reihe verschiedener *Wörterbuchtypen*. Beispiele sind Rechtschreibwörterbücher, (zwei- oder mehrsprachige) Übersetzungswörterbücher, Häufigkeitswörterbücher, grammatische (z.B. morphologische) Wörterbücher, Synonym-Wörterbücher (die Wörter gleicher oder ähnlicher Bedeutung angeben), etymologische (oder Herkunfts-)Wörterbücher u.v.m.
- Die **Lexikologie** untersucht Strukturen und Zusammenhänge im Wortbestand. Auch hier ist zu beobachten, dass die beiden Termini Lexikographie und Lexikologie in der Praxis oftmals nicht unterschieden werden.

8.1 Wortschatzgröße

Fragen nach dem Umfang des Wortschatzes einer Sprache beschäftigen immer wieder Laien, die Medien und populärwissenschaftliche Veröffentlichungen, aber auch Fachwissenschaftler verschiedener Richtungen. In diesem Zusammenhang werden oft konkrete Zahlen genannt; manchmal hört man auch Aussagen wie „Die englische Sprache ist die mit dem größten Wortschatz“. Wie verhält es sich nun tatsächlich?

Generell sollte man solche Aussagen mit größter Vorsicht behandeln; denn meist kommen sie aufgrund willkürlicher Entscheidungen und ohne eine Grundlage zustande, die die Vergleichbarkeit verschiedener Aussagen ermöglichen würde.

Jede Aussage über die Größe eines Wortbestands hängt davon ab, wie er zuvor definiert worden ist. Wörterbücher können immer nur einen Ausschnitt wiedergeben, der von den betreffenden Lexikographen im Hinblick auf die geplante Verwendung und den Adressatenkreis festgelegt wird. In keinem Fall kann ein Wörterbuch vollständig sein, denn noch während daran gearbeitet wird, verändert sich der Wortbestand laufend.

Das größte lexikographische Projekt für das Deutsche begannen die Sprachwissenschaftler Jacob und Wilhelm Grimm im Jahre 1838 mit über 80 Mitarbeitern. Sie veranschlagten damals sechs Bände, was sich als Fehlschätzung herausstellte. Nach ihrem Tod übernahmen andere ihre Arbeit, bis schließlich 123 Jahre später im Jahr 1961 der 32. und letzte Band mit 350.000 Stichwörtern erscheinen konnte.

Außerdem ist zu bedenken, dass die Festlegung des lexikalischen Ausschnitts, der betrachtet wird, auch davon abhängt, inwieweit *regionale Besonderheiten* und vor allem *Fachwortbestände* berücksichtigt werden sollen. Das größte moderne Wörterbuch des Deutschen (*Duden. Das Große Wörterbuch der deutschen Sprache*, 10 Bände, ³1999) umfasst mehr als 200.000 Stichwörter mit über 300.000 Bedeutungsangaben. Berücksichtigt man auch Fachwortschätze, so kommt man auf Millionen von Wörtern.

¹ Anmerkung: Im allgemeinen Sprachgebrauch gilt: Ein Wörterbuch macht Angaben zu Wörtern, während ein Lexikon über Sachverhalte Auskunft gibt. Die Grenze zwischen beiden kann fließend sein.

Leichter ist der Umgang mit dem Vokabular eines Texts; denn hierbei handelt es sich um eine abgeschlossene Menge, die man durch Abzählen erfassen kann. Allerdings gibt es auch hier Hürden: Vor der Zählung muss man sich überlegen, wie man die in einem Text vorkommenden Wortformen zählen will. Soll man z.B. in dem Satz „er reiste ab“ 2 oder 3 Wörter zählen? Einerseits könnte man argumentieren, „abreisen“ sei ein Wort, andererseits können wir deutlich drei Wörter ausmachen; denn zwischen „reiste“ und „ab“ können weitere Wörter wie z.B. „leider“ oder „gestern nach einem ausgiebigen Frühstück“ stehen. Dieses und andere Probleme lassen sich nur im Hinblick auf den Zweck der jeweiligen Untersuchung bzw. auf die Eigenschaften lösen, um die es bei der Zählung geht.

Schwieriger ist es wieder, die Größe individueller Vokabulare von Einzelpersonen zu bestimmen. Man unterscheidet zwischen dem aktiven und dem passiven Wortschatz, also zwischen den Wörtern, die eine Person selbst verwendet, und denen, die sie lediglich versteht. Schätzungen belaufen sich auf einen Umfang von einigen Hundert Wörtern bis zu mehr als 20.000 – je nach Bildungsstand und anderen Umständen. Durchschnittlich wird mit 6.000 bis 10.000 Wörtern gerechnet.

Ein großer Teil der Probleme im Zusammenhang mit Fragen nach dem Wortbestand ist eine Konsequenz aus der Problematik des *Wortbegriffs*.

8.2 Wortbegriff

Ähnlich wie die Silbe gilt auch das Wort als intuitiv klarer Begriff. Versuche einer linguistisch befriedigenden Definition sind aber wiederum von willkürlich gesetzten Kriterien abhängig, oder sie leiden an mangelnder Widerspruchsfreiheit. Mögliche Kriterien sind (1) das **phonetische** (Wort als durch eindeutigen Akzent und künstlich gesprochene Pausen identifizierbare Einheit), (2) das in der Computerlinguistik nicht selten verwendete, aber linguistisch sehr problematische **orthographische** (Wort als Buchstabenfolge zwischen Leerzeichen und anderen Trennsymbolen), (3) das **morphologische** Wort (Einheit, die der Flexion unterliegt und durch Wortbildungsregeln gekennzeichnet ist), (4) das **syntaktische** (kleinste auf Satzebene austauschbare und einschiebbare Einheit) und (5) das **semantische** Wort (kleinster selbstständiger Träger einer Bedeutung).

In jedem Fall ist es wichtig, zwischen **Wortform** und **Lexem** zu unterscheiden. Wörter begegnen uns im Text als Wortformen. Ein z.B. lexikalisch-semantisch oder morphologisch bestimmtes Wort wie „Kind“ finden wir (in vielen Sprachen) in mehreren solcher Wortformen vor: Kind, Kindes, Kinder, Kindern. In agglutinierenden Sprachen können Wörter Dutzende von verschiedenen Wortformen aufweisen, in flektierenden Sprachen sind es weniger, meist in Abhängigkeit von der Wortart. Die Grundbedeutung der verschiedenen Formen (die „lexikalische“ Bedeutung) ist in jedem Fall dieselbe, Unterschiede bestehen lediglich in den grammatischen Kategorien. Abstrahiert man von diesen Unterschieden und Flexionsformen, spricht man von dem **Lexem** „Kind“. Man kann auch sagen, dass das Lexem alle Wortformen eines Worts zusammenfasst, ähnlich wie ein Morph als die Menge von Allomorphen gesehen werden kann, die eine gemeinsame Bedeutung besitzen.

Vertreten wird ein Lexem durch einen kanonischen Stellvertreter, unter dem es auch im Wörterbuch zu finden ist. Im Deutschen besteht die Konvention, Substantive durch die Wortform des Nominativs im Singular vertreten zu lassen, Verben durch den Infinitiv, Adjektive durch die prädikative Form. In anderen Sprachen werden auch andere Formen gewählt, Verben werden z.B. im Ungarischen durch die 3. Person Singular Präsens vertreten, im klassischen Hebräischen durch die 3. Person Singular Maskulinum *Qal* Perfekt, im Lateinischen und im Klassischen Griechischen durch die 1. Person Singular Indikativ Aktiv. Den Stellvertreter selbst nennt man **Lemma**. Das Lemma ist sozusagen der Name des jeweiligen Lexems.

Ein weiteres Begriffspaar, das nicht nur im Bereich der Lexik von Bedeutung ist, soll ebenfalls an dieser Stelle eingeführt werden: *Type* und *Token*. Unter einem Type versteht man in Bezug auf Wörter die verschiedenen Wörter, die in einem Text vorkommen. Im folgenden Textausschnitt sind die Types nummeriert:

Eine(1) Apfelsine(2) und(3) eine(1) Zitrone(4) stehen(5) hier(6) für(7) die(8) Zitrusfrüchte(9). Für(7) die(8) anderen(10) Obstsorten(11) steht(12) ein(13) Apfel(14).

Das zweite Vorkommen von *eine* ist kein neuer Type; denn dieses Wort kam schon vor. Dasselbe gilt für das zweite Vorkommen von *für*; es wurde nicht mit einer eigenen Nummer markiert, weil der Type schon verwendet wurde. Die verschiedenen Vorkommen gehören zum selben Type. Diese einzelnen Vorkommen der Wörter nennt man Token. Der kleine Textausschnitt besteht also aus 14 Types mit insgesamt 17 Token.

In diesem Beispiel wurden *Wortformentypes* markiert. Verschiedene Wortformen desselben Lexems wurden also eigenen Types zugeordnet. Je nach Untersuchungsziel kann man auch *Lexemtypes* betrachten, sodass verschiedene Wortformen zu Tokens desselben *Lexemtypes* werden. Das Beispiel von oben sähe dann so aus:

Eine(1) Apfelsine(2) und(3) eine(1) Zitrone(4) stehen(5) hier(6) für(7) die(8) Zitrusfrüchte(9). Für(7) die(8) anderen(10) Obstsorten(11) steht(5) ein(1) Apfel(12).

8.3 Lemmatisierung

Eine wichtige Prozedur in der Computerlinguistik, die Voraussetzung für viele weitergehende Aufgaben der Sprachanalyse und -verarbeitung ist, besteht darin, zu einer gegebenen Wortform das zugehörige Lemma herauszufinden, um es z.B. in einem (elektronischen) Wörterbuch nachschlagen zu können. Diese Prozedur nennt man **Lemmatisierung**.

Um die Lemmatisierung automatisch-maschinell durchzuführen, gibt es grundsätzlich zwei Methoden. Eine Methode ist das Nachschlagen der zu lemmatisierenden Wortform in einem **Vollformenwörterbuch**, in dem jede einzelne flektierte Form jedes Wort einen Eintrag besitzt und auf das jeweils korrekte Lemma (bzw. die in Frage kommenden Lemmata, s.u.) verweist. Diese Methode ist sehr speicheraufwändig, aber schnell und unkompliziert.

Die zweite Methode beruht auf möglichst vollständigen Listen aller (Allo-)Morphe und einem **Regelwerk**. Die zu analysierende Wortform wird zunächst in potenzielle Allomorphe segmentiert. *Ratte* z.B. würde dabei in *rat* und *-te* (Suffix 3. Pers. Prät.) zerlegt, weil an dieser Stelle noch nicht bekannt ist, dass die Form *Ratte* nicht als Verbform von *raten* in Frage kommt. Die Regeln, die in diesem Schritt anzuwenden sind, müssen dafür sorgen, dass z.B. *Bein* nicht in das Präfix *be-* und das Wort *in* zerlegt wird, oder solche Segmentierungen müssen vor dem nächsten Schritt ausgefiltert werden (im Fall des Verbs *beinhalten* allerdings wäre die Segmentierung *be-in-halt-en* korrekt). Dann müssen die Flexionsmorphe separiert werden. Entweder wird der Rest dann in einem Stamm-Lexikon nachgeschlagen oder es werden die Flexionsmorphe hinzugefügt, die für die Lemmata in Frage kommen, und in einem üblichen Wortverzeichnis gesucht.

Anhand der bei der Wortform gefundenen Flexionsmorphe lässt sich die Wortform anschließend mit einer vollständigen morphologischen Information versehen.

Meist gibt es mehrere, oft sehr viele prinzipiell mögliche Segmentierungen. Das Verfahren muss dafür sorgen, dass nur solche Segmentierungen in Betracht gezogen werden, die den Regeln der Morphologie (Morphotaktik) der betreffenden Sprache genügen und bei Zusammensetzung der Wortform ein gültiges, restloses Ergebnis erbringen. So wäre die Segmentierung *Nach-t* nicht zulässig, weil es kein Verb **nachen* gibt und weil die Präposition *nach* nicht mit einem Verbalsuffix wie *-t* verbunden werden darf. Die morphologische Methode ist

weniger speicheraufwändig als die auf einem Vollformenwörterbuch basierende, aber sehr komplex und schwer zu warten.

Auch das beste Verfahren kann oft keine eindeutige Analyse liefern. Einige Probleme sind:

- Eigennamenerkennung: (1) Man benötigt zur Erkennung von Eigennamen ein entsprechendes Verzeichnis, das jedoch immer unvollständig sein muss. (2) Viele Eigennamen sind formgleich mit anderen Wörtern (*Koch, Kohl, Köhler ...*);

- im Wörterbuch nicht gefundene Wörter;

- Abkürzungen;

- Komposita: (1) In Sprachen wie dem Deutschen oder dem Ungarischen sind unbegrenzte Bildungen von Komposita möglich. Mit einem aufwändigen Zerlegungsverfahren ist dieses Problem allerdings lösbar. (2) In Sprachen wie dem Englischen oder Französischen (durch Nachahmung neuerdings auch in vielen unkorrigierten deutschen Texten) werden Komposita in der Regel nicht zusammen geschrieben. Dass ein Kompositum vorliegt, kann dann im Rahmen der Lemmatisierung nicht ohne Weiteres erkannt werden;

Ambiguitäten: [die] *Macht* / *Mach-t* [sie das?]; *Stau-becken* / *Staub-ecken*; alle Arten von Homonymie (s.u. im Abschnitt „Lexikalische Relationen“).

Mildern lassen sich viele Probleme durch Berücksichtigung der Umgebung des zu lemmatisierenden Worts.

8.4 Lexikalische Semantik

Bei der Betrachtung der Bedeutung von Wörtern unterscheidet man zwischen der denotativen und der konnotativen Bedeutung. Während die denotative Bedeutung intersubjektiv ist (d.h. alle Sprecher der Sprache stimmen mehr oder weniger darin überein, worin die jeweilige Bedeutung besteht), bezeichnet man mit konnotativ die subjektiven oder privaten Bedeutungsanteile. Deutlich wird das, wenn man vergleicht, was z.B. ein Wort wie *Bruder* oder *Schwester* für verschiedene Menschen bedeutet, je nach dem, ob sie selbst einen oder mehrere Geschwister haben, vielleicht verloren haben, sich wünschen, wie das Verhältnis ist oder war usw.

Die denotative Bedeutung kann auf zwei verschiedene Arten angegeben werden: (1) extensional durch Aufzählen aller Objekte, die darunter fallen (z.B. für *Monat* die 12 Monate von Januar bis Dezember). Dies ist natürlich nur in bestimmten Fällen möglich – wenn nämlich die Anzahl dieser Objekte endlich ist. (2) intensional durch Angabe der Eigenschaften, die ein Objekt besitzen muss, um unter die Bedeutung zu fallen (z.B. für *Monat* „eine Zeiteinheit von 28 bis 31 aufeinanderfolgenden Tagen, in die ein Jahr eingeteilt wird“). Dies ist auch im Falle unendlich vieler Objekte möglich.

Wichtig ist die Tatsache, dass sich die Bedeutung von Wörtern prinzipiell von der Bedeutung von Sätzen unterscheidet. Letztere besteht in einer Aussage (der man auch einen Wahrheitswert wie *wahr* oder *falsch* zuweisen kann). Die *Wortbedeutung* besteht nicht in einer Aussage, sondern in einem Verweis auf ein Einzelobjekt (bei Eigennamen), eine Klasse von Objekten (wie bei *Apfel* oder *Katze*), auf eine Relation (wie bei *Vater*; was eine bestimmte Relation zwischen zwei Lebewesen bezeichnet) usw.

In der „**Sinnrelationalen Semantik**“ unterscheidet man einige fundamentale Relationen zwischen Wörtern. Die wichtigsten sind:

Synonymie: Synonyme sind Wörter gleicher Bedeutung (strikte Synonymie, z.B. *Apfelsine* – *Orange*; *Tatsache* – *Fakt*, *grundsätzlich* – *prinzipiell*) bzw. ähnlicher Bedeutung (partielle Synonymie, z.B. *zeitgemäß* – *modern*). Zu beachten ist, dass auch strikte Synonyme nicht in allen Kontexten durch einander ersetzbar sind, da sie oft stilistisch oder regional markiert sind.

Synonymie trifft man nicht nur in der Lexik an, sondern auch in anderen linguistischen Bereichen. So gibt es z.B. synonyme Morphe und syntaktische Konstruktionen.

Antonymie: Das Antonym zu einem Wort bezeichnet das Gegenteil, den Gegenpol o.ä. Bsp.: *hoch – tief; lang – kurz; über – unter*.

Homonymie: Homonyme sind Wörter gleicher Form aber verschiedener Bedeutung. Viele Sprachwissenschaftler bezeichnen nur solche Wortpaare als homonym, die unterschiedlicher Herkunft sind (z.B. *Tau*: Seil/Niederschlag/griechischer Buchstabe). Man unterscheidet zwei Formen von Homonymie: Homophonie und Homographie. Homophon sind Wörter, die gleich klingen (auch wenn sie unterschiedlich geschrieben werden) wie *sieben*: ‚Zahl/Tätigkeit mit einem Sieb‘. Als homograph bezeichnet man Wörter gleicher Schreibung (auch wenn sie unterschiedlich gesprochen werden wie *modern*: ‚zeitgemäß/zerfallen durch Mikroorganismen‘. Auch Homonymie existiert auf mehreren linguistischen Analyseebenen.

Polysemie: Ähnlich wie Homonyme sind Polyseme formgleiche Wörter verschiedener Bedeutung. Allerdings geht man hier von einer Bedeutungsverwandtschaft aus (durch Bedeutungsverschiebung, Spezialisierung, Übertragung entstanden) wie bei *Bank*: ‚Möbelstück/Geldinstitut‘, *Flügel*: ‚Körperteil/Musikinstrument/Gebäudeteil‘.

Polylexie: Nur in der quantitativen Linguistik verwendet. Bezeichnet allgemein jede lexikalische Mehrdeutigkeit; der Unterschied zwischen Polysemie und Homonymie spielt dabei keine Rolle.

Hyponymie / Hyperonymie: Ein Hyponym ist ein Unterbegriff. So ist *Stuhl* ein Hyponym von *Möbel*; umgekehrt ist *Möbel* ein Hyperonym (Oberbegriff) von *Stuhl*. *Möbel* hat mehrere Hyponyme, die zueinander **Kohyponyme** sind: *Stuhl, Tisch, Bank, Bett ...*
Aus der Tatsache, dass *Tisch* ein Hyponym von *Möbel* ist und dass *Schreibtisch* ein Hyponym von *Tisch* ist, kann man logisch korrekt schließen, dass *Schreibtisch* ein *Möbel* ist (Transitivität der Relation).

Meronymie / Holonymie: Dieser Terminus bezeichnet die Teil-Ganzes-Relation (partitive Relation). Ein *Rad* ist ein Teil eines *Fahrzeugs*, die *Nabe* Teil des *Rads*. Die umgekehrte Relation nennt man Holonymie.

Lexikalische Funktionen: (Apresjan/Mel'čuk: Smysl'-text-teorija [Bedeutung-Text-Theorie])

Um eine semantische Beziehung zwischen zwei Wörtern auszudrücken, müssen in Abhängigkeit vom Bezugswort jeweils ‚passende‘ Wörter verwendet werden, auch wenn es immer um dieselbe semantische Beziehung geht: Wenn jemand viel raucht, sagt man, er sei ein *starker Raucher*. Wer viel isst, wird dagegen *guter Esser* genannt. Ähnlich heißt es *schwerer Trinker, ergiebige Niederschläge, reiche Ausbeute, herbe/bittere Enttäuschung, schallende Ohrfeige, starke Säure, intensives Blau, kräftiger Druck, hohes Risiko, scharfer/starker Kontrast* etc. und allen anderen semantischen Beziehungen, wobei oft Alternativen bestehen (*fester* oder *tiefer Schlaf*). Diese Tatsache besteht unabhängig von den Wortarten der beteiligten Wörter. Beispiele für Verb-Substantiv-Syntagmen: Man *stellt* eine *Frage*, *gibt* eine *Antwort*, *macht* eine *Aussage*, *hält* eine *Rede*.

Welche Wörter dabei zusammenpassen, ist in jeder Sprache gesondert geregelt. Ein starker Raucher heißt im Englischen nicht *strong* oder *powerful*, sondern *heavy smoker*. Vor allem durch schlechte Übersetzungen englischsprachiger Texte (nicht zuletzt im Bereich der Computertechnik) und Filme werden zunehmend englische Regeln von deutschsprachigen Spre-

chern übernommen (*brandneu* statt *nagelneu*, willkommen *zu* statt richtig willkommen *bei* etc.)

Als Hilfe für korrekte Übersetzungen vor allem auch in computerlinguistischen Systemen fasst man gleiche semantische Beziehungen (lexikalische Funktionen) zusammen und gibt für jede, z.B. in einer Datenbank, die verwendbaren Wortpaare an. Beispiele:

Etwas existiert:

Vertrag, Hoffnung, Bedenken – bestehen;

Frieden, Krieg, Freude, Ruhe – herrschen;

Gerücht, Plage – umgehen;

Gefahr – bestehen, drohen

...

Etwas geht zu Ende:

Freude, Leid, Hoffnung, Leidenschaft – vergehen/erlöschen;

Bedenken, Hoffnung – schwinden;

Vorstellung – enden;

Vorrat – zu Neige gehen/sich erschöpfen

...

8.5 Sprichwörter und Phraseologismen

In den Bereich der Lexik gehören auch Mehrwort-Ausdrücke, wenn ihre Bedeutung nicht aus den Bedeutungen der Bestandteile und grammatischen Regeln geschlossen werden kann (nicht-kompositionale Semantik). Es ist sinnvoll, dabei Sprichwörter und Phraseologismen zu unterscheiden.

Spruchwörter werden im Rahmen der **Parömiologie** untersucht. Sie bestehen formal aus einem vollständigen Satz, sie drücken eine meist alte, allgemeine Lehre oder Weisheit aus (*Wer anderen eine Grube gräbt, fällt selbst hinein*), die über die unmittelbare Bedeutung des Satzes hinausgeht. Sprichwörter lassen sich nicht direkt übersetzen; oft aber gibt es in anderen Sprachen Sprichwörter gleicher oder ähnlicher Bedeutung, aber dies ist in hohem Grade von der Verwandtschaft der Kulturkreise abhängig.

Allerdings gibt es auch in diesem Bereich die Übernahme schlechter Übersetzungen, vor allem aus dem Englischen. Es scheint, dass heute unter den jüngeren Menschen das deutsche Sprichwort *Morgenstund hat Gold im Mund* kaum noch bekannt ist; dagegen ist die Kenntnis der wörtlichen Übersetzung von *The early bird catches the worm* verbreitet und wird für ein deutsches Sprichwort gehalten.

Phraseologismen sind formelhafte Ausdrücke, die im Gegensatz zu Sprichwörtern keinen Satz bilden. Ebenfalls im Unterschied zu Sprichwörtern sind Phraseologismen veränderbar (Flexion der einzelnen Wörter, Einschübe und Verschiebungen) und richten sich nach dem grammatischen Kontext. Es gibt verschiedene Arten von Phraseologismen:

- Formeln (Gruß- und Anredeformeln, Flüche, Segen und andere magische Formeln etc.);
- Zwillingsformeln wie *Kind und Kegel*, *Bausch und Bogen*, *Pauken und Trompeten*, *sage und schreibe*, *[durch] dick und dünn*, *[bei] Nacht und Nebel* ...
- Redewendungen wie *ins Gras beißen* (für *sterben*), *mit Zitronen handeln* (für *erfolglose Geschäfte betreiben*) oder *das Kind mit dem Bade ausschütten* (für *radikale Mittel einsetzen, die sich letztlich negativ auswirken*);

- Redensarten (zitatartige, fest erstarrte Wendungen) wie *Ende gut – alles gut, mitgefangen – mitgehangen, strenge Rechnung, gute Freundschaft* oder *harte Schale, weicher Kern*.

8.6 Quantitative Lexikographie und Lexikologie

Alle Eigenschaften und Relationen von lexikalischen Einheiten werden auch mit quantitativen Mitteln untersucht. Das ist eine so gewaltige Zahl von Untersuchungsgegenständen, dass sie nicht sinnvoll in diesem Rahmen dargestellt werden können. Die Verteilung der Häufigkeit von Wörtern (die sich auch in Häufigkeitswörterbüchern findet), der Länge, Polysemie und vieles mehr, Abhängigkeiten zwischen Eigenschaften (je häufiger ein Wort, desto kürzer; je länger, desto weniger Bedeutungen und desto speziellere Bedeutungen), dynamische Eigenschaften von Texten wie der Verlauf des Type-Token-Verhältnisses usw., werden als mathematische Funktionen beschrieben.

Zur quantitativen Linguistik gibt es ein gesondertes Skript.