Ressourcenbeitrag: Katharsis

Marie Flüh [1]

1. Universität Hamburg

forTEXT

Thema:	Sentimentanalyse	DOI:	10.48694/fortext.3795
Jahrgang:	1	Ausgabe:	7
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2019-09-02 auf fortext.net
Lizenz:	CC BY SA		open ∂ access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Kurzbeschreibung

Katharsis ist ein webbasiertes Werkzeug bzw. eine Ressource für die quantitative Dramenanalyse von mehr als 100 deutschsprachigen Dramen (vgl. Korpus). Die Ergebnisse der Analyse werden als interaktive Webschnittstelle dargestellt. Darüber hinaus stehen zwölf Dramen Gotthold Ephraim Lessings für eine Sentimentanalyse (Flüh 2024) zur Verfügung.

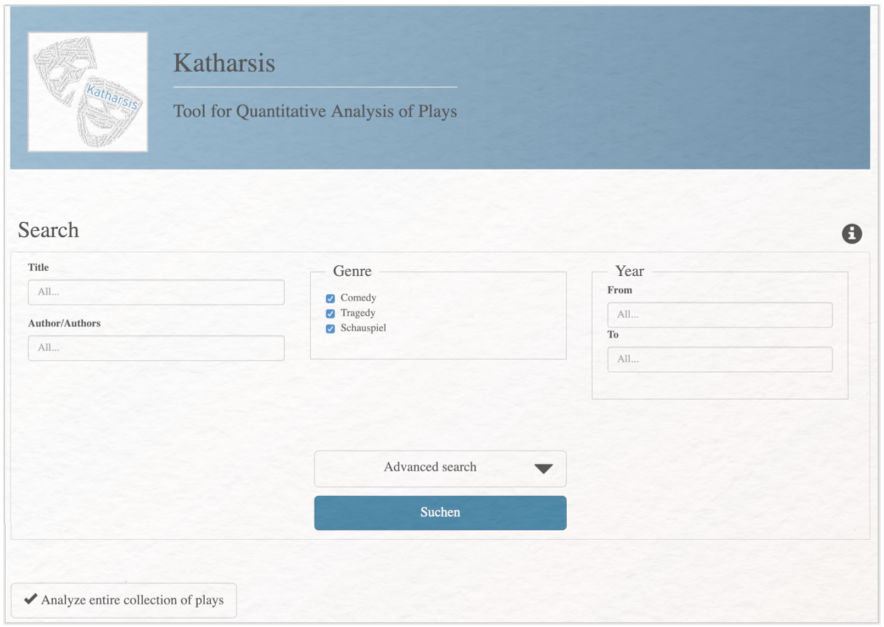

Abb. 1: Die Startseite von Katharsis

Steckbrief

- http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/index.html
- Werkzeug zur quantitativen Dramenanalyse (TextGrid Repository (Horstmann 2024)) von 177 deutschsprachigen Dramen (Burghardt u. a. 2018) und Sentimentanalyse ausgewählter Dramen Gotthold Ephraim Lessings (Schmidt 2017)
- Beinahe ausschließlich kanonisierte Dramen (Schauspiele, Trauerspiele oder Komödien) aus dem Zeitraum zwischen den Jahren 1646 bis 1838 mit einem Schwerpunkt auf Veröffentlichungen aus dem Jahr 1755
- Individuelle Auswahl eines oder mehrerer Dramen für die Einzelanalyse oder vergleichende Analysen
- Quantitative Parameter der Einzelanalysen: Anzahl der Akte und Figuren, Konfigurationsdichte (Populationsdichte innerhalb des Dramas), Berechnung der Häufigkeitsverteilung zu Repliken und Replikenlänge; Darstellung der Ergebnisse u. a. als interaktive Figurenkonfigurationsmatrix
- Parameter der Sentimentanalyse: Berechnung der Sentiments im gesamten Drama (Sentiment-Metriken: Polarität oder Emotionen); einbezogene Emotionen: Zorn, Erwartung, Ekel, Angst, Freude, Traurigkeit,

Überraschung und Vertrauen; interaktive Kreisdiagramme zu: Sentiment-Anteil im Drama, Sentiment-Anteile pro Akt, Sentiment-Anteile pro Szene; interaktive Verlaufsdiagramme zu Sentiment im Drama pro Akt, Sentiment in Szene pro Akt, Szenen im Dramenverlauf, Repliken
- Methodische Tradition und theoretischer Hintergrund: Quantitative Methoden in Literatur- und Kulturwissenschaft, Distant Reading (Moretti 2013), Dramen als frühe Belege einer mathematischen Poetik (Marcus 1973)
- Metadaten: Titel, Autor*in, Genre, Datum der Erstveröffentlichung, Redeanteile

2. Anwendungsbeispiel

Sie möchten Konfigurationen in *Lady Johanna Gray* (1757) von Christoph Martin Wieland untersuchen und anschließend mit weiteren Dramen aus dem 18. Jahrhundert vergleichen.

3. Diskussion

3.1 Kann ich Katharsis für wissenschaftliche Arbeiten nutzen?

Ja, erste Fallstudien wurden erfolgreich durchgeführt (Dennerlein 2015).

3.2 Wie benutzerfreundlich ist die Arbeit mit Katharsis?

Der übersichtliche und strukturierte Aufbau der Webseite (vgl. GUI) ermöglicht einen unmittelbaren und unkomplizierten Einstieg in die quantitative Dramenanalyse, gleichzeitig werden zitierbare Analyseergebnisse in ansprechend aufbereiteter Form präsentiert: Durch die Verbindung aus Medieninformatik und Literaturwissenschaft entsteht hier ein großer Mehrwert. Zusätzlich hilft der „Info“-Button bei der Bedienung von Katharsis. Er ist allerdings nur auf der Startseite vorhanden (siehe Abb. 1). Es steht kein Handbuch zur Verfügung. Zu den einzelnen Parametern der quantitativen Dramenanalyse finden Sie unter „overview“ eine kurze Einführung, die bei der Interpretation der Visualisierungen hilfreich ist. Eine Erläuterung der durch die Sentimentanalyse entstehenden Daten bzw. Angaben zur Grundlage der Berechnungen fehlen. Sentiment- und Dramenanalyse sind auf das vorgegebene Korpus beschränkt. Die Analyseergebnisse können Sie als PNG-Datei herunterladen.

4. Wie funktioniert die Textsuche in Katharsis?

Katharsis funktioniert zuverlässig, ist ausschließlich online verfügbar und lässt sich mit jedem Browser verwenden. Die integrierte Sentimentanalyse-Funktion befindet sich zum jetzigen Zeitpunkt in der Beta-Version.
Grundsätzlich gilt: Sie können entweder eine Einzelanalyse eines Dramas aus dem Korpus durchführen oder mehrere Dramen vergleichend analysieren lassen. Sie beginnen die quantitative Dramenanalyse, indem Sie die gewünschten Dramen in die Suchmaske (vgl. Query) eingeben oder indem Sie über den „Search“-Button aus einer Liste des gesamten Korpus diejenigen Dramen markieren, die Sie in eine vergleichende Analyse inkludieren möchten. In der Voreinstellung sind alle 117 Dramen markiert. Entfernen Sie das Häkchen in der ersten Reihe der Übersicht, um diese Vorauswahl zu deaktivieren und eine eigene Auswahl zu treffen.
Durch einen Klick auf „Analyze entire collection of plays“ bestätigen Sie Ihre Auswahl und starten die Analyse.
Bei der Interpretation der Ergebnisse sollte der theoretische Hintergrund berücksichtigt werden: Die mathematische Dramenanalyse basiert im Kern auf der Berechnung von Figurenkonstellationen (Konfigurationen). Die Konfigurationsmatrix stellt das Auftreten aller Figuren in sämtlichen Szenen zusammenfassend dar (siehe Abb. 2).

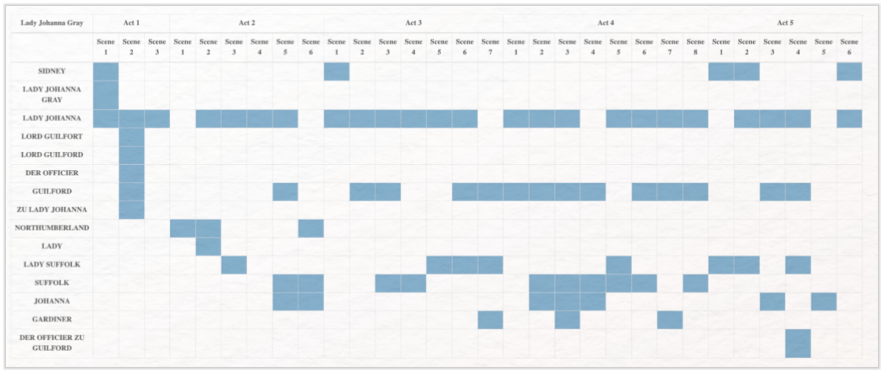

Abb. 2: Im Rahmen einer Einzelanalyse erstellte Konfigurationsmatrix von Christoph Martin Wielands Lady Johanna Gray

Detaillierte Angaben finden Sie, indem Sie innerhalb der Konfigurationsmatrix mit dem Cursor über die jeweiligen Figuren hovern (siehe Abb. 3).

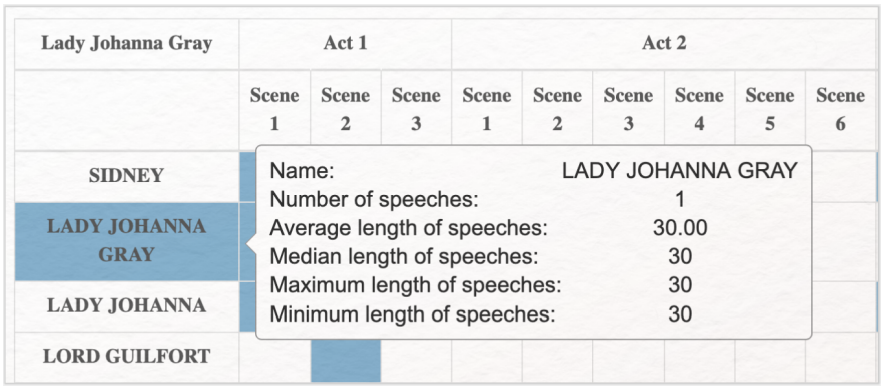

Abb. 3: Detailansicht der Konfigurationsmatrix: Analyse der Redeanteile Lady Johanna Grays

Über die Menüleiste navigieren Sie in weitere Analysemodi: *Configuration matrix*, *Structural analysis* (siehe Abb. 4), *Speaker analysis* (siehe Abb. 5), *Distribution of speech* oder *Sentiment Analysis*.

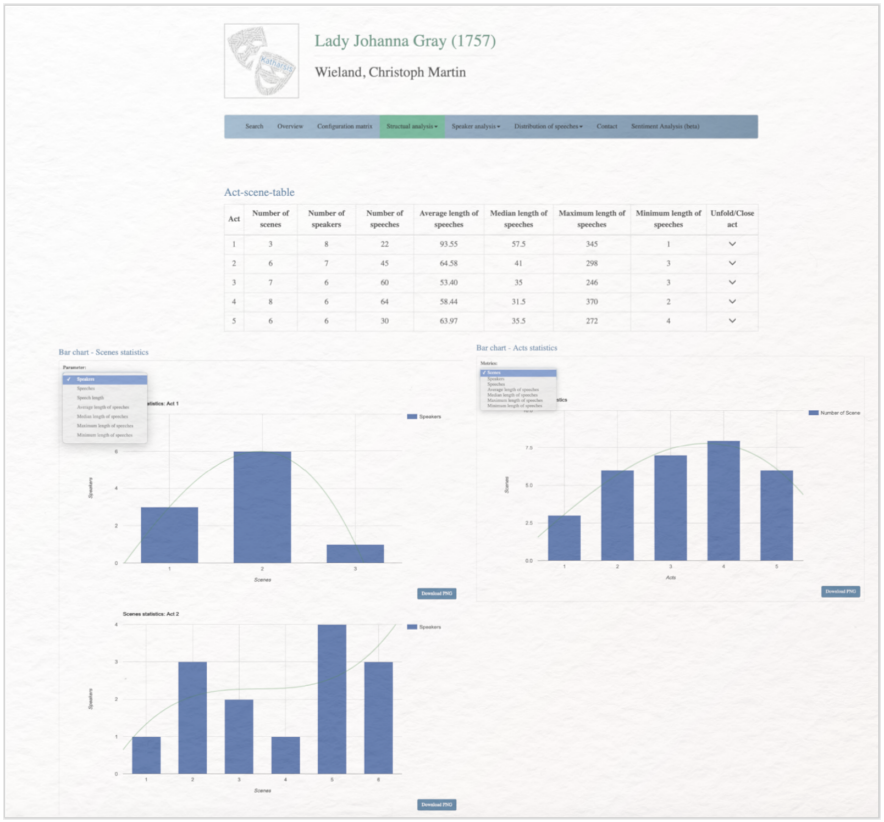

Abb. 4: Strukturanalyse von Wielands Lady Johanna Gray: Gesamtüberblick sowie Analyse aller Szenen und Akte

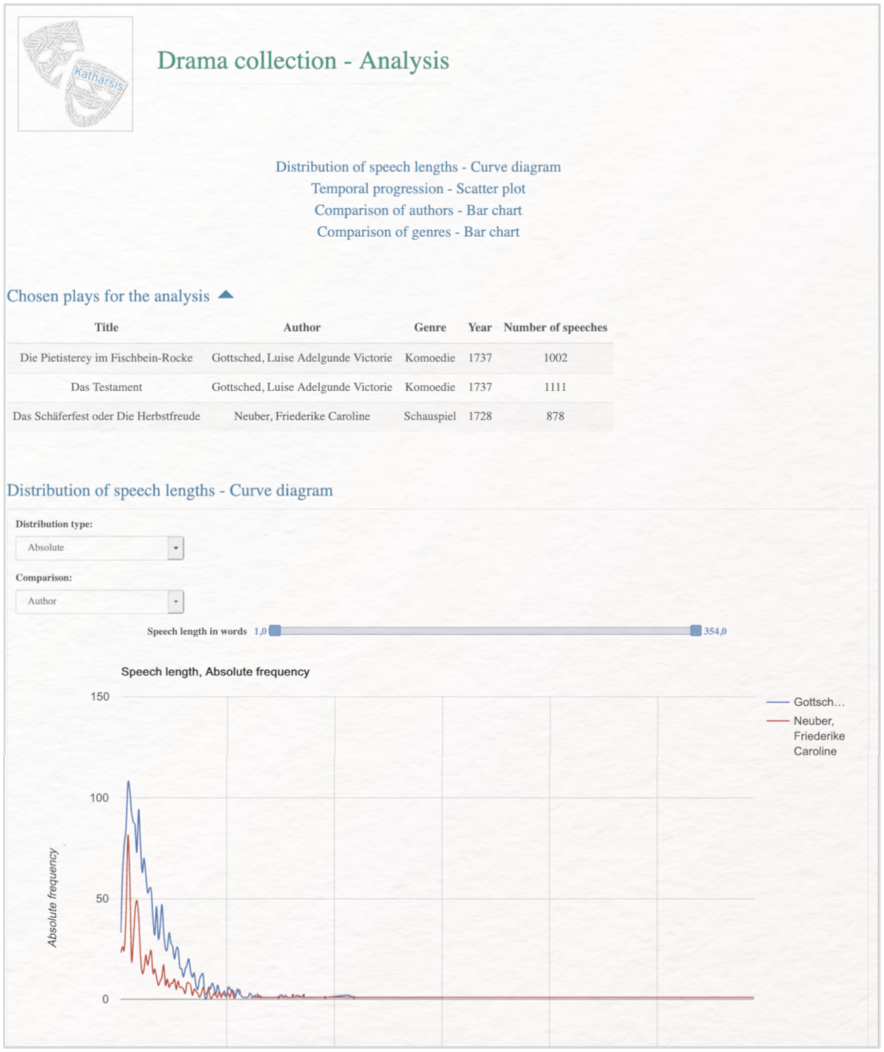

Abb. 5: Vergleichende Quantitative Analyse der drei Dramen im Korpus, die von Frauen verfasst wurden: Kurvendiagramm der Redeanteile

Externe und weiterführende Links

- Katharsis: http://lauchblatt.github.io/QuantitativeDramenanalyseDH2015/index.html (Letzter Zugriff: 04.09.2024)

Bibliographie

Burghardt, Manuel, Katrin Dennerlein, Thomas Schmidt, Johanna Mühlenfeld und Christian Wolff. 2018. Katharsis - Ein Werkzeug für die quantitative Dramenanalyse. In: Hamburg. doi: 10.5283/epub.37582, (zugegriffen: 26. August 2019).

Dennerlein, Katrin. 2015. Measuring the average population densities of plays. A case study of Andreas Gryphius, Christian Weise and Gotthold Ephraim Lessing. *Semicerchio. Rivista di poesia comparata LIII*: 80–88.

Flüh, Marie. 2024. Methodenbeitrag: Sentimentanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3797, https://fortext.net/routinen/methoden/sentimentanalyse.

Horstmann, Jan. 2024. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3794, https://fortext.net/ressourcen/textsammlungen/textgrid-repository.

Marcus, Solomon. 1973. *Mathematische Poetik*. Frankfurt am Main: Athenäum.

Moretti, Franco. 2013. *Distant Reading*. London, New York: Verso.

Schmidt, Thomas. 2017. Gefühl ist alles; Name ist Schall und Rauch - Der Einsatz von Sentiment Analysis in der quantitativen Dramenanalyse. Masterarbeit im Fach Medieninformatik am Institut für Information und Medien, Sprache und Kultur. Regensburg: Universität Regensburg.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

POS PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.