Prof. Dr. H. Schweizer

Eberhard Karls Universität, Tübingen

PHRASEOLOGIE-ONLINE

Die Thesen bilden den Hintergrund des mündlichen Beitrags. Zugleich stellen sie ein erstes README für das vorzustellende Computerprogramm dar. Dort eingeschlossen sind weitere Erläuterungen und Hilfetexte.

1. Phraseologie ist – wie jedes Sprachphänomen – auf drei Ebenen zu betrachten: (a) auf der der Ausdrucksseite (nachfolgend: Syntax genannt – den Unterschied zwischen dem informatischen und dem linguistischen „Syntax"-Verständnis beachten!), (b) auf der der wörtlichen Bedeutung (Semantik) und (c) auf der der Pragmatik.
2. Der Impuls, überhaupt von *Phrasem* zu reden, liegt auf der Ebene der Ausdrucks-Syntax: Wörter (*tokens*) bilden Ketten (*strings*).
3. Ist die Wahrscheinlichkeit ihrer Nachbarschaft hoch, redet man von *Phrasem, Idiom, geprägter Wendung, Formel, Sprichwort* o.ä.
4. Solche Wortnachbarschaften auffallender Häufigkeit aufzudecken, ist der Computer natürlich das ideale, weil präzise und schnelle Werkzeug.
5. Mit der Befunderhebung auf ausdruckssyntaktischer Ebene ist erst der **phraseologische Impuls** gefunden, der Sprachbefund, der Anlass zu weiteren Analysen, dann auf semantischer und pragmatischer Ebene, gibt.
6. Für die sprachstatistische Befunderhebung bieten wir in Tübingen das Programm *CoMOn (=Corpus Matching Online)* an: *http://www-ct.informatik.uni-tuebingen.de/Comon/www/*. Das Programm dient hier zur Illustration eines methodisch-algorithmischen Vorgehens.
7. Wenn eine Fragestellung leicht und schlüssig programmierbar ist, liegt darin eine zusätzliche Bestätigung für die Richtigkeit der zugrundeliegenden Theorie. Unter letzterer wird das semiotische Grundfaktum jeglicher Sprache verstanden: die klare und **eindeutige Trennung von Ausdruck und Bedeutung**. Dass „Bedeutung" noch weiter zu differenzieren ist – „wörtlich" oder „indirekt" –,

betrifft nicht die Basisunterscheidung von Ausdruck und Bedeutung.

8. Der Rechner, indem man ihn auf der *Ausdrucksebene* arbeiten lässt, kann seine hohe Effizienz ausspielen.

9. Die unzweideutige Fokussierung auf die Ausdrücke (*strings*) korrigiert zugleich eine Inkonsequenz, die traditionelle *Konkordanzen* in Buchform meist bieten: es handelt sich dort eben nicht lediglich um reine Wortlisten (u. U. mit Kontextangabe). Häufig sind doch noch Kriterien aus dem Bedeutungsbereich in die Anordnung des Wortmaterials eingeflossen.

10. Um mit *CoMOn* zu arbeiten, müssen dort die entsprechenden Textkorpora integriert werden. Das verlangt vorbereitende Entscheidungen, etwa die – im Blick auf deutsche Texte –, dass alle Satzzeichen eliminiert werden, dass alle Großschreibungen in Kleinschreibungen verwandelt werden. Ein und dieselbe Wortform, die einmal am Satzanfang steht, ansonsten im Satz (und dann klein geschrieben) würde vom Computer nicht als identisch erkannt.

11. Zur Vorbereitung gehört auch, dass eine Zählung integriert wird. Die Kapitelgliederung etwa eines Romans gibt selbst schon eine Struktur vor. Bisweilen ist im Original die Zählung schon vorgesehen. Wenn nicht, wird sie von uns ergänzt. Dann geht es auch darum, dass die Paragrafen innerhalb eines Kapitels als eigene, vom Autor gewollte optische Substrukturen berücksichtigt werden. Daher zählen wir innerhalb eines Kapitels die Paragrafen durch.

12. Ein Korpus kann somit – analog zur biblischen Zählung – gezählt werden. „BUCHKKK,PPP; “ Dabei ist „BUCH” entweder das gesamte Korpus oder – wo es sich anbietet – ein Buch im Rahmen des Korpus (wie bei der Bibel). „BUCH” selbst ist unterteilt nach „Kapiteln” = „KKK” (= “dreistellige Kapitelnummer”). Innerhalb der „Kapitel” werden „Paragrafen” = „PPP” jeweils mit „001” beginnend durchgezählt. Die „Paragrafen” wurden in lockerer Analogie zu biblischen „Versen” gewählt.

13. Aktuell bietet *CoMOn* folgende Korpora an: hebräisches Altes Testament, griechisches Altes Testament, griechisches Neues Testament, Koran arabisch, Koran deutsch, G. Grass „Die Blechtrommel”, NEGRA-Korpus. –

Erweiterungen sind möglich.

14. **Praktische Recherche:** Durch Auswahl des Korpus und Bestimmung eines Suchtextes darin (via „BUCHKKK,PPP;" – Anfangs-, dann Endposition) bestimmt der Nutzer, zu welchem *Suchtext* er phraseologische Ergebnisse gewinnen möchte. Aus Speichergründen unterliegt die Länge des Suchtextes Beschränkungen.

15. Vor dem Start des Programms ist noch zu klären, welche *Mindestlänge* von Suchtreffern gewünscht wird. Voreingestellt ist die Länge „3". Das kann nach unten wie nach oben korrigiert werden. Eine Maximallänge muss nicht eingegeben werden, da der Algorithmus selbst feststellt, bis zu welcher Länge ein *string* des Suchtextes im übrigen Korpus eine Entsprechung findet. Der Benutzer kann sich also überraschen lassen.

16. Die Standardsuche zielt darauf, *identische* Entsprechungen zu finden. Wir arbeiten daran, dass auch *Auslassungen, Umstellungen, Ähnlichkeiten* gefunden werden können. Der Rechenaufwand ist dann deutlich höher.

17. **Phraseologie** basiert aber wesentlich auf ***Identitäten***. Die stilistischen Effekte, die auf Ausdrucksseite geweckt werden können, verflüchtigen sich, je weniger fest eine Wortkette verwendet wird.

18. *CoMOn* durchläuft also das gesamte Korpus und nimmt die *strings* des Suchtextes zum Maßstab: immer, wenn es im Korpus eine Übereinstimmung zu einer Wortkette aus dem Suchtext gibt, wird diese Übereinstimmung als Treffer gemerkt.

19. Nach dem Durchlauf wird zuerst ein tabellarisches Ergebnis geboten, das darüber informiert, wie viele Treffer gefunden wurden, welches ihre Durchschnittslänge ist, wie lange die Suche gedauert hatte.

20. Die Suchdauer spielt sich im Sekundenbereich ab. Dagegen – eigene Erinnerung – dauerte ein sorgfältiges Durchprüfen einer biblischen Erzählung mittels Buch-Konkordanz einmal 14 Tage... Von den oben erwähnten Nachteilen einer Buchkonkordanz und der Fehleranfälligkeit durch Ermüdung usw. ganz abgesehen.

21. **Ergebnisdarstellung:** Über „*generate conclusion*” lässt sich das Suchergebnis übersichtlich anzeigen: Die in der Vertikalen zweigeteilte Tabelle erinnert zunächst daran, unter welchen Bedingungen die Suche durchgeführt worden war. Alle Einstellungen und auch der definierte Suchtext werden aufgeführt.

22. Die zweite Tabelle führt in der linken Spalte den **kompletten Suchtext** auf. Gibt es zu einem *string* keinen Treffer im restlichen Korpus, wird dies durch „---” in der zweiten Spalte angezeigt. Fand sich ein Treffer, wird in der ersten Spalte im Klartext der *string*, um den es aktuell geht, wiedergegeben, in der zweiten Spalte dessen Länge vermerkt. In der dritten folgen die Stellenangaben. In der vierten die Anzahl der Stellen mit genau diesem Treffer.

23. Im Überblick lässt sich damit leicht feststellen, wo der Suchtext mit dem restlichen Korpus auf *Wortketten-Ebene* verbunden ist, wo dagegen nicht. Häufen sich „---”, so spricht dies für eine Passage im Suchtext, die kreativ und eigenständig formuliert worden war. Kommen in anderen Textbereichen Treffer mit vielen Belegen vor, so klinkt sich diese Passage in gängigen, allgemein üblichen Sprachgebrauch ein. Nach den stilistischen Effekten und Motiven kann, ja muss, im einen wie im andern Fall gefragt werden.

24. Zu beachten ist die Möglichkeit, dass ***eine*** **Wortkette in abgestufter Form mehrfach** aufgeführt ist: Zunächst werden die längsten Entsprechungen genannt. Danach die kürzeren. Zu einer 8er-Kette des Suchtextes kann es vereinzelte Entsprechungen geben. Aber auch zu einer 6er-Kette daraus, und zusätzlich zu einer 3er-Kette daraus. Derartige geschachtelte Befunde werden übersichtlich separat genannt.

25. In der Demo werde ich mich auf Befunde aus der „Blechtrommel” von Günter Grass beschränken. Nur angedeutet sei, dass natürlich die *grafische* Wiedergabe der anderen genannten Korpora eigene Herausforderungen stellte. Es ging um die griechische, hebräische und arabische Schrift, dabei um Akzente, Vokalzeichen, die unterschiedliche Form eines einzelnen Buchstabens je nach Position im Text. Und schließlich – in den letzten beiden Fällen – um die andere

Schreibrichtung. – Hierbei waren Kompromisslösungen notwendig. Das Ergebnis ist jeweils doch so, dass es – wie wir finden – sich sehen lassen kann.

26. **Danksagung:** *CoMOn* basiert – informatisch gesehen – auf einer Diplomarbeit von *Michael Pach*, sowie deren Anpassung und Integrierung in ein Java-Applet durch *Serhiy Bykh.*

27. **Theorierahmen:** Sprachverwendung geschieht, weil „Bedeutungen" vermittelt werden sollen. „Bedeutungen" lassen sich aber nicht transportieren. Sie werden im kognitiven System des Menschen je neu (re-)konstruiert, auf der Basis des Wissens, das dort schon vorhanden ist.

28. Zum Transport der Stimuli dienen materielle „Vehikel" (optisch = Schriftzeichen, akustisch = Laute, haptisch = Gebärden). Sie sind es, die die kognitive Eigenaktivität der Rezipienten auslösen. Sie sind es aber auch, die sehr häufig übersehen, unterbewertet und in der Analyse vernachlässigt werden.

29. Semiotiker haben schon das „Körper / Leib-Geist"-Bild verwendet, um das Verhältnis von „Ausdruck – Bedeutung" zu charakterisieren. Folglich käme eine linguistische Fixierung auf die „Bedeutungen" einer „Leibverachtung" gleich, wäre ein „linguistischer Manichäismus". Konzentrierte Ausdrucksanalyse – wie vorgestellt – würde dagegen der Ausdrucksebene neu zu ihrem Recht verhelfen.

30. „Was ist das schwerste von allem? – Was dir das leichteste dünket: Mit den Augen zu sehn, was vor den Augen dir liegt" (J. W. v. Goethe, XENION).