



Introduction

by Dr Achim Beutner

1. Introduction

di-lemmata offers more than the usual digital presentation of literary texts. It enables greater in-depth analysis of the texts and gives researchers the opportunity to direct the pattern and direction of their enquiry on an individual basis.

The principal parts of the program are:

- texts easily accessible to linguistic analysis
- lemmatised word lists
- a flexible concordance utility package
- comparison of word lists

di-lemmata opens completely new possibilities for computer-assisted text analysis. It is a set of complex tools and utilities for exploring literary texts, regardless of whether the users' interests are of a personal or professional nature. It can provide them with extensive, verifiable results. **di-lemmata** is a comprehensive utility package: the program supplies the data, the conclusions based on the meaning and interpretation of the texts is, naturally, the responsibility of the user; perhaps the information resulting from the program leads to further research and a deeper understanding and enjoyment of the material under study.

1.1 The Texts

Currently, the following works are available online:

Gottfried August Bürger (Gedichte)

Johann Wolfgang von Goethe (Gedichte letzter Hand; West-östlicher Divan)

Friedrich von Schiller (Gedichte)

Friedrich Hölderlin (Gedichte)

Ludwig Uhland (Gedichte)

Joseph von Eichendorff (Gedichte und Epen)

August Graf von Platen (Gedichte)

Annette von Droste-Hülshoff (Gedichte)

Heinrich Heine (Gedichte)

Nikolas Lenau (sämtliche Gedichte)

Eduard Mörike (sämtliche Gedichte)

Friedrich Hebbel (sämtliche Gedichte)

Gottfried Keller (Gedichte)

Conrad Ferdinand Meyer (sämtliche Gedichte)

Hugo von Hofmannsthal (Gedichte)

Rainer Maria Rilke (Gedichte)

Georg Trakl (sämtliche Werke)

Georg Heym (sämtliche Werke)

Alfred Lichtenstein (sämtliche Werke)

Further texts are in preparation and will be available at a later date.

The objective is to compile a representative corpus of modern German literature enabling a wider and more comprehensive analysis of the material.

A fundamental difference to many of the available sources of digitally prepared texts is that the texts presented here have been linguistically developed and prepared: i.e. each corpus or part thereof (contained in separate folders) is accompanied by a lemmatised word list or dictionary. The approach to literary criticism and the posing of relevant questions concerning poetic vocabulary and usage ought to be based on texts subjected to this type of preparation. This would include information on the distribution and frequency of the principal word classes (cf. 1.2), comparisons of the vocabulary used by different authors and detailed examination of the contexts etc. (A simple index of words is of less value for this type of research.)

As the corpus increases in size and variety, the possibilities for more extensive analysis are assured. This may lead to a deeper understanding of aspects of change in theme and subject matter in modern German literature.

1.2 The Word Classes

The determination of word classes in **di-lemmata** is based on the considerations outlined below.

Every attempt to classify word forms brings with it the difficulty that the morphological, syntactical and semantic aspects are intermingled and lead to differing results. A unified and generally accepted classification does not exist and according to linguists will not exist in future. Consequently, one could have used the systematic found in the latest Grammatik-Duden as a basis for the process of lemmatisation. For **di-lemmata** a different approach has been selected. The classification used here (after the words in the texts have been lemmatised) does not attempt to render a detailed grammatical description, rather is subject to the main objective of the program viz. to allow and open new possibilities of computer-assisted text analysis. In other words, this classification is not an attempt to propose a new grammatical model, but simply to allow a framework for new enquiries.

At this point, questions as to sense and meaning, (Sinn und Bedeutung, Signifikant und Signifikat, Systematik und Pragmatik etc.) are not under discussion (nor indeed of particular relevance to **di-lemmata**). There is undoubtedly some justification in asserting that the semantic content of a text lies mainly with the principal parts of speech such as noun, adjective and verb, and any analysis of a poet's vocabulary and use of words is primarily concentrated on these specific forms.

They are differentiated from the remaining word classes (referred to here as "Restklasse" = "Remainder") by their "semantic function" as well as being "open classes". These so-called "open classes" include lexemes that arise new; others become disused and disappear in the course of time. Forms such as articles, pronouns and conjunctions form the "Restklasse", they are considered "closed classes" as few, if any, new lexemes are formed.

Their number is mostly constant.

Between the principal parts of speech and the "Restklasse" is a substantial quantitative difference. This is especially noticeable with an increased volume of texts. "Restklasse" word forms have a relatively constant number of lexemes whereas the principal parts of speech tend to vary in quantity. The opposite is true of the frequency with which lexemes occur in the texts: forms in the "Restklasse" such as articles and conjunctions ("und") are much more frequent. The only exception to this is the verb "sein" which also occurs in its function of an auxiliary. In Trakl's work, for example, no principal parts of speech appear in the top 10 most frequent lexemes, only 2 nouns and 3 adjectives when the top 25 are taken into consideration. They are: "Nacht" and "Schatten" and "dunkel", "schwarz" and "blau".

The "Restklasse" is not of outstanding (semantic) importance in questions of literary criticism. Nonetheless, an attempt was made to divide it into sub-classes – somewhat limited and with some exceptions. Every unambiguous lexeme was sorted to the appropriate word class, but a differentiation as found in formal grammars was not undertaken. All pronouns were classified together without further differentiation as were also the various forms of conjunctions.

Following the overall objective and intention of the project – viz. meaningful support of literary analysis – the word forms proper names, foreign words and quotations are listed under C. It is understood that these 3 categories do not strictly belong here; nonetheless it was deemed helpful to treat the 3 groups in this fashion.

Finally, the scope of the lemmatisation process should be mentioned: all the poetic texts were lemmatised. Diaries, letters and other writings (not considered poetic texts) were not lemmatised. A Register of Persons has been compiled for these texts.

The Word Classes:

- | | |
|--------------------------|-----------------------|
| A. Main Parts of Speech: | 1. Noun |
| | 2. Adjective |
| | 3. Verb |
| B. Die Restklasse: | 4. Adverb |
| | 5. Article |
| | 6. Pronouns |
| | 7. Pronominal Adverbs |
| | 8. Interrogatives |
| | 9. Prepositions |
| | 10. Conjunctions |
| | 11. Ambiguous Entries |
| C. | 12. Proper Names |
| | 13. Foreign Words |
| | 14. Quotations |

2. The Program

The three main parts of *di-lemmata* are:

- the library (the corpora of texts)
- the word lists
- the application program

2.1 The Library

The library contains all the works currently collected in the program, sorted into folders and sub-folders. There is a biography of each author at the beginning of the folder.

The generally available editions of the authors' works are used as a basis for the project. These tend to use the normalised orthography of the time when the edition was printed. A representation using a critical-historical edition (if one were available) was avoided on the following grounds:

a) Many editions are in themselves inconsistent (e.g. the Weimarer Edition of Goethe's work). This would entail editing the text or being left with many written variations and lemmata which, despite reflecting the spellings, would be of no value to the dictionary (especially from a semantic point of view).

b) 18th and 19th Century texts tend to spell 't' as 'th', 20th Century texts will write simply 't'. It is easily seen that this would result in recording a lot of doubles (e.g. "That" versus "Tat") where no semantic difference exists.

Subsuming the old spelling under a normalised lemma, though possible, would entail an enormous amount of extra work. This stands in no relation to any value that could be gained from a differentiated representation. In consequence, this course of action has not been followed. This has been restricted to normalised editions. In the case of Goethe's or Hölderlin's work, this is unfortunate, but perhaps a solution to this dilemma will be found in future.

2.2 The Word Lists

The complete vocabulary of every text included in the program is available in the word lists. The content of a list depends on which folder or text has been selected. For example, if Goethe has been selected then every entry from the Goethe corpus in *di-lemmata* will appear in the word list. If a single folder is chosen then only the vocabulary from the texts in this folder will appear in the word list.

The word list will contain details on every lemma, lexeme, word class and frequency reflecting the chosen corpus. The list can be sorted by word stem, word class or frequency. By clicking on an example, the reference and the surrounding context of a word can be displayed.

There are a number of options and filters which enable the user to determine exactly

what information should be included in any particular word list. For example, one could compile a list of a specific part of speech or several different ones or even all of them. Lexemes can be marked in one word list and these entries can be copied to other specially created word lists.

It is possible to devise lists according to morphological markers enabling analysis of word formation patterns such as all the forms that contain prefixes or suffixes, e.g. "Er...ung". Using these relatively simple means, a systematic and extensive analysis of a lemmatised vocabulary can be achieved.

2.3 Concordances

One of the most interesting aspects of examining literary texts is the question of how lexemes are distributed in sentences and phrases. In particular, how does their use in literature differ from what could be expected to occur in general, non-literary use of the written language.

In order to examine the environment of a word or group of words, text analysis has relied on computers to produce various types of concordances such as KWIC index (key word in context index). *di-lemmata* offers the user a broad palette of tools and utilities.

The user can (via the word list) select one or more lexemes or even a complete list as a key on which to base the analysis. Using filters it is possible to further specify environment and content that will be produced around the key word or phrase. One could request that a context is produced only when a noun is preceded by an adjective or even where only a specific adjective appears. The user is able to determine what, if anything, precedes or follows the key word (or phrase whether this be several types of word forms or lexemes). The result of this search produces a concordance tailored to the user's requirements and available for more intensive study.

2.4 Comparisons

One of the principal features of this program is its ability to enable comparisons of vocabulary between various corpora. This allows the user to compare: i) the complete works of various authors ii) various collections of texts from a particular author iii) individual texts from one author as compared to various other authors. The amount of vocabulary under study can be determined by the user: a study of the total vocabulary, several word classes (only the main parts of speech), or restricted to one e.g. nouns. Over and above this, there are options which allow the user to further refine the degree of comparison to be undertaken.

The value of vocabulary comparisons within one source can be demonstrated by examining the relatively small corpus left by Trakl. So even a comparison of vocabulary contained in the only two publications which occurred during Trakl's lifetime - the collections "Gedichte" und "Sebastian im Traum" can yield useful results. These texts are both roughly the same size but exhibit a tendency to diverge in both theme and style. (There follows an example: cf. 3. below). From this it is obvious that these tools can provide

greater results when applied to the works of someone as prolific as Goethe. Tools of comparison as envisaged here may perhaps be able to provide data on the difference in his early works and those of later life. Up till now this has only be achieved on the basis of single words or phrases.

If the corpus is extended sufficiently, a multitude of possibilities to examine and compare authors, and even epochs, will ensue.

3. Example: The Use of Adjectives of Colour in the Works of Trakl

The following example uses the collections in Trakl's "Gedichte" and "Sebastian im Traum". To preserve clarity, only the ten most frequently used adjectives from each are used in the comparison.


Wortstamm	Klasse	Trakl, Gedichte	Rang	Trakl, Sebastian im Traum	Rang
dunkel	Adj			61	1
leise	Adj	24	3	58	2
blau	Adj	23	4	53	3
schwarz	Adj	39	1	51	4
purpurn	Adj			37	5
still	Adj			32	6
silbern	Adj			30	7
grün	Adj			29	8
rot	Adj	17	8	25	9
sanft	Adj	23	5	25	10
braun	Adj	28	2		
weiß	Adj	23	6		
alt	Adj	20	7		
golden	Adj	16	9		
schön	Adj	16	10		

Examining this list, it is apparent that the adjectives "dunkel", "purpurn", "silbern", "still" and "grün" grow in importance in the author's later work ("Sebastian im Traum") whereas others such as "leise", "blau", "schwarz", appear to remain constant.

Using these results and support from a concordance program to examine other word and context lists, one might gain some insight into the "poet's workshop".

It is obvious to every reader that adjectives of colour play an important part in Trakl's work. This has been widely discussed in academic circles. Trakl's use of "blau", "rot" is reminiscent of parallels in the world of expressionism.

The following table shows the quantitative development in Trakl's use of adjectives of colour.



Wortstamm	Klasse	Gedichte	Sebastian im Traum	Sammlung 1909
rot	Adj	17	25	5
grün	Adj	2	29	3
blau	Adj	23	53	2
weiß	Adj	23	21	2
schwarz	Adj	39	51	1
grau	Adj	12	5	1
braun	Adj	28	12	
golden	Adj	16	19	
gelb	Adj	10	6	
silbern	Adj	6	30	
purpurn	Adj	5	37	

The proportion of adjectives of colour in the total number of adjectives:

	Adjektive insgesamt	Farbadjektive	%
Sammlung 1909	471	14	2,97
Gedichte	902	181	20,07
Sebastian im Traum	1211	288	23,78

There is an obvious difference between the occurrence of adjectives of colour in the early work (Sammlung 1909) and later poems. In a letter to a friend he distanced himself from his early work and talked of his "heiß errungenen Manier" (struggle), of how he employed a technique using "vier Strophenzeilen, vier einzelne Bildteile zu einem einzigen Eindruck" (four lines, four separate parts of a picture are moulded into one single impression). This is known in literary criticism as (expressionist) "Reihenstil". An independent observer might conclude that it may have had less to do with the "heiß errungene Manier" and use of the so-called Reihenstil, than the change of vocabulary he used. (This does not only apply to the adjectives cited here, but to the vocabulary in total, which has been shown to be the case using the tools available in **di-lemmata**.)

In the later work "Sebastian im Traum", one could go as far as to claim that no noun could appear without being in danger of an adjective of colour accompanying it. So in the collection of poems in "Gedichte" and in "Sebastian im Traum", (the editions published with Trakl's approval), there is at least one adjective of colour in every poem. One could describe this as a constituent element of Trakl's later lyrics.

In order to illustrate the use of adjectives of colour further, a list can be drawn up showing the possible combinations of adjectives of colour and nouns.

1. Concrete Nouns

These nouns depend on perceptible reality and environment. The following distinctions have been made:

- a) Objects that have been manufactured, e.g. das grüne Kleid (the green dress) or

die rote Tischlampe (the red table lamp)

These objects can be combined with any adjective of colour (which may well stretch the imagination, but nonetheless remain within the bounds of possibility).

b) Natural Objects, e.g. der weiße Strand (the white beach).

Natural objects can be characterised by those colours in which they normally occur which limits or rather restricts the descriptive possibilities. Despite this, there are areas where these possibilities are expanded in a variety of ways and used in situations which one could describe as unrealistic combinations. These are generally referred to as metaphors. An example of the "impossible" combination of adjectives of colour with concrete nouns can be found in Paul Celan's famous "schwarze Milch" (black milk).

2. Abstract Nouns

a) In everyday speech, adjectives of colour can be found in combination with abstract nouns in stylised forms ("verblaßten Metaphern", faded metaphors) such as "graue Theorie" (unproven, dull theory), "blauer Montag" (Monday off work especially after a "rough" weekend). Quite often expressions such as these are predetermined by the symbolic value of the adjectives of colour, especially by "schwarz" and "weiß" (black and white), e.g. "weißer Sonntag" (White Sunday), "schwarze Gedanken" (evil / morbid thoughts). It can be noted that "schwarz" and "weiß", contain an inherent symbolism which can not easily be extracted from the cultural environment in which they are used. So for example "gelb" (yellow) symbolises hope in the USA. Colours can be used symbolically in everyday speech such as when politicians talk about a "grüne Zukunft" (environmentally green future).

b) Every combination of adjectives of colour with abstract nouns other than those in a) are categorised as "symbols". The term "symbol" has been used to avoid the use of "kühnen Metapher" ("bold metaphor") Every metaphor is based on analogy, but when that analogy is not readily identifiable, one can no longer use the term metaphor.

It is plain that this classification is not unambiguous as it is not always possible to differentiate between the concrete and abstract nouns as the "lexeme" (without context) may belong to both groups. An example of this is the "blaue Blume" of the Romantic era (often found in Trakl's work). This has long since developed into a symbol in which both parts point to their origins, but taken together must be considered as an abstract.

In order to remain within the limits of this introduction, the study has been confined to the distribution of the adjective "blau". This adjective of colour appears as an attribute in combination with the following nouns:

1. Concrete Nouns	
a) Manufactured	b) Natural
Band Bilder Gewand (2) Glocken Kahn Kugeln Mantel (4) Panzer (Kriegers) Räumen Schleiern Tabernakel Zimmern	Antlitz Augen (4) Bach Blume (10) Brauen (2) Falter Farben (2) Firne (3) Fluß (3) Früchte Gebirge Grunde Höhle (2) Kalvarienhügel Kristall Kuppeln (Himmelweiten) Lider (3) Luft (2) Quell (7) Schatten (2) Schleim Seen Tauben Teich (2) Tier (2) Wasser (5) Wild (6) Woge Wolke
2. Abstract Nouns	
a) symbolic, conventional	b) Metaphor, Symbol
Glanz Frühling	Abend (6) Augenblick Gestalt (Menschen) Klage Kühle Lachen (2) Nacht Odem Orgelgeleier Seele (2) Stille (4) Stimme Ton

Trakl combines 56 different lexemes with the adjective "blau", of which 19 (in bold script) appear to be used metaphorically, or perhaps phrased more cautiously, in an unconven-

tional fashion. This simple example serves as a warning as to how one interprets **the** meaning of the adjective of colour "blau" in this context. A number of doubts arise.

A closer examination of two passages may help here:

- a) O wie stille ein Gang den *blauen Fluß* hinab
Vergessenes sinnend, da im grünen Geäst
Die Drossel ein Fremdes in den Untergang rief.
- b) Jener aber ging die steinernen Stufen des Mönchsbergs hinab,
Ein *blaues Lächeln* im Antlitz und seltsam verpuppt
In seine stillere Kindheit und starb;

It is not really adventurous to maintain that the "blaue Fluß" in example a) will be understood differently by the reader from the "blaue Lächeln" in b). Nothing is more conventional or natural (aside from any environmental issues) than a "blauer Fluß" (blue river). In fact, there is nothing exceptional here that would induce the reader to stop short, reflect deeply about it and express any form of astonishment or surprise. Something quite different can be expected when the reader comes across "blaue Lächeln". It is not so much that this is a most unusual phrasing (many expressions we take for granted nowadays were originally "unusual"). A modern "professional" reader of lyric poetry would not experience them as such, rather tend to expect new "unusual" phrasing and formulation which expand and enrich their view of the world. How then can one take the "blauen Lächeln" and discover a connection between "blau" and "Lächeln"? Goethe's famous "graue Theorie" (dull theory) in opposition to the "grünen Baum des Lebens" (green tree of life - vibrant, active life) is instantly perceived as a metaphor because the analogies can easily be derived. Things are different in Trakl's "blauem Lächeln"; there is nothing analogous to be found in one's own area of experience or language. So, as the reader readily accepts the combination of "blau" and "Fluß" imperceptibly on the basis of his own experience and language, the reader views Trakl's "blaues Lächeln" as a contradiction that can not be resolved in his experience and language spectrum. The reader is vexed and irritated. On the premise that the author consciously chose his words (many would argue this point), the reader now needs to find an explanation or reason for this choice. This is the point, in our opinion, where every interpretation of aesthetic works begins and also the proliferation of the various readings and interpretations known as "Lesarten-Pluralismus". One result of the "irritation" experienced by coming across a "kühnen Metapher" or "Chiffre" in Trakl is the conclusion that the literary critic or expert finds the adjective "blau" (in our example) has a special place and value in Trakl's work.. This is a matter of conjecture and requires a more detailed analysis which can be achieved quickly and thoroughly using the reference basis available in **di-lemmata**.

From this initial study of the use of the adjective "blau", it is obvious that many other relevant questions can be posed. For example, which other attributes a noun may exhibit other than blue. The following list shows results using the noun "Schatten" (shadow / shade):

linker Kontext ▾	Schlüsselwort	rechter Kontext
zarten	Schatten	
weißer	Schatten	
weiße	Schatten	
verzerrte	Schatten	
verlorner	Schatten	
trauervolle	Schatten	
toter	Schatten	
teuren	Schatten	
stillen	Schatten	
Schwere	Schatten	
schwarzer	Schatten	
schwarzen	Schatten	
schwarzen	Schatten	
schwarze	Schatten	
schwarze	Schatten	
roter	Schatten	
rosiger	Schatten	
rosiger	Schatten	
phantastische	Schatten	
Nächtigen	Schatten	
nächtigen	Schatten	
Nächtigen	Schatten	
mondverschlungenen	Schatten	
letzter	Schatten	
jammervoller	Schatten	
heraufbeschworener	Schatten	
grünen	Schatten	
grünen	Schatten	
gramvollen	Schatten	
goldenen	Schatten	
gewaltiger	Schatten	
furchtbaren	Schatten	
friedlosen	Schatten	
feuchten	Schatten	
feuchten	Schatten	
ernsten	Schatten	
brauner	Schatten	
braunen	Schatten	
bläulicher	Schatten	
Bläuliche	Schatten	
Bläuliche	Schatten	
blauen	Schatten	
blauen	Schatten	
blaue	Schatten	

Once again, it is evident that aside from the conventional combinations such as "schwarze Schatten" (on the same semantic level as the "blaue Fluß"), there are very unusual, almost erratic attributes to be found, such as "mondverschlungene" (moon-entwined),

"gramvolle" (melancholy) Schatten. An extended analysis of nominal phrases in Trakl's work could be achieved on the basis of references gained here.

4. Future Direction

"Aller Anfang ist schwer", is a popular saying. Hesse reminds us that "Allem Anfang wohnt ein Zauber inne". The gospel of St. John in German translation begins "Im Anfang war das Wort". In our opinion **di-lemmata** has made a new beginning, a new way to approach the subject of computer-assisted literary text analysis.

The example used in the previous chapter demonstrates some of the possibilities opened up to users of **di-lemmata**. It represents only a very small selection of possible uses. More extensive analyses are planned or currently being undertaken e.g. the vocabulary used in Goethe's poems.

The main difficulty does not lie in obtaining the texts, but in their organisation and administration. There is always the danger that one won't be able to see the wood for the trees - i.e. to lose one's way in a labyrinth of statistics, word lists etc. Completeness and accuracy are worthy objectives, but clarity and manageable presentation are more important. This has resulted in a division of the material such as in the example shown of Trakl's use of the adjective "blau".

Every user of **di-lemmata** is cordially invited to begin their own analysis and research into the vocabulary of the authors currently contained in the corpus.

We should be delighted to receive your articles, suggestions for improvements, notification of any errors and last, not least, your critical appraisal.