Collocations in Lexicography

Ulrich Heid

Wien:
Workshop on
Computational Approaches to Collocations,
23 August 2002

Overview

- Definitional elements from lexicographic theory
  - Two lines of tradition:
    * Statistically oriented research
    * Didactically oriented research
  - Towards an integrated view
- Collocations vs. other multiword items:
  Collocations – selectional classes – idioms
- Lexicographic tasks wrt. collocations
  - A descriptive programme
  - Detailed description of collocations
  - Presentational aspects: the user view
- Corpus lexicographic tasks wrt. collocations:
  The case of dictionary updating

Collocation – Definitional elements from linguistic and lexicographical theory

Introduction

Collocation research in Lexicography

"... is a mess"

(Herbert Ernst Wiegand, in recent a discussion)

- Definitions of linguistic terminology are sometimes vague
- Lexicographic definitions overlap with those used in other areas of applied (computational) linguistics, e.g. Information Retrieval, IE, terminology extraction, etc.
- Different approaches and lines of tradition in general language lexicography
- Not (yet) enough data and theorizing lexicography for specialized languages and terminology

Lines of tradition in lexicography

In line with [Klotz 2000]

- Statistically oriented line:
  Statistically significant cooccurrence of items
- Didactically oriented line:
  Items relevant for the production of "naturally sounding" language

Collocation – Definitional elements from linguistic and lexicographical theory

The statistically oriented line of research (1/2)

British Contextualism

- The famous citation: [Firth 1957]:
  "You shall know a word by the company it keeps"
- [Halliday 1966: 152]:
  " [...] lexis seems to require the recognition merely of linear co-occurrence together with some measure of signficant proximity, either a scale, or at least a cut off point"
- [Sinclair 1991]: *Corpus, Concordance, Collocation*
- [Kjellmer 1994]: *A dictionary of English collocations*

Examples from [Sinclair 1991]: 118-120:
The word *back*

- *be back at school; drive back down to ...; come back from Paris; follow him back into the wood*
- *arrive back on Tuesday; bring it back into fashion; it dates back to 1950*
- *we had to go back to the hotel ; he went back home; go back to bed; I' ve just got back from the office*

Collocation – Definitional elements from linguistic and lexicographical theory

The statistically oriented line of research (2/2)

British Contextualism: Major definitional elements
Cf. [Klotz 2000: 72s.]

- Cooccurrence of two elements:
  Originally of arbitrary word class (Sinclair, Kjellmer)
- Statistically significant cooccurrence
- Preferably within a certain distance:
  "Collocational Span" ("Window size" ...)
- Observation:
  Groups of collocations with morphologically related words:
  - [Halliday 1966]:
    strong argument – argue strongly
    – strength of an argument – strengthen an argument
  - [Altenberg/Eeg-Olofsson 1990]:
    drink heavily – heavy drinker – heavy drinking
  - ◦ Over-interpreted:
    Collocations independent from word classes
  - ◦ But [Greenbaum 1988: 114]:
    desperately needed – desperate need
    *badly needed – *bad need*

Collocation – Definitional elements from linguistic and lexicographical theory

The didactically oriented line of research (1/3)

(Pedagogical) Lexicography

- [Hausmann 1979, 1985] etc.:
  Distinction between *base* and *collocate*:
  - The collocate is lexically selected by the base:
    "Ein Partner determiniert, der andere wird determiniert" ([Hausmann 1985])
  - Collocation selection is relevant for text production:
    Speaking about the base, look for adequate collocate
  - Collocation selection must be learned
- Meaning of a word used as a collocate (e.g. *pay attention*) differs from its meaning outside the collocation (Hausmann, [Benson 1985], etc.)
- Attempts to classify the lexical semantic properties of collocates in collocations:
  Mel'čuk:
  Lexical Functions, Explanatory Combinatorial Dictionary
- Most recently:
  [Runcie (Ed.) 2002]:
  Oxford Collocations Dictionary for Students of English

Collocation – Definitional elements from linguistic and lexicographical theory

The didactically oriented line of research (2/3)

(Pedagogical) Lexicography: Major definitional elements

- Combination of two lexemes (plus closed-class items)
- Word class of elements is relevant:
  Collocations follow the rules of syntax
- Statistical significance vs. usuality:
  HAUSMANN: "Kombinationen von auffallender Üblichkeit":
  Usuality corresponds to typicality (significance):
  " [...] typische, spezifische und charakteristische Zweierkombinationen von Wörtern"

Examples from [HAUSMANN 1985]:

eingefleischter Junggeselle	`‘confirmed bachelor’`
schütteres Haar	`‘thinning hair’`
heikles Thema	`‘delicate subject’`
Geld abheben	`‘draw money’`
Unfall bauen	`‘cause an accident’`
Rechnung begleichen	`‘settle a bill’`

Collocation – Definitional elements from linguistic and lexicographical theory

The didactically oriented line of research (3/3)

<u>Contrastive relevance of collocations</u>

- Collocate selection in the target language:
  - `*Empfang* + *machen*` → *Empfang geben*
  - `*Zusammenkunft* + *machen*` → *Zusammenkunft abhalten*
- Collocates need not be "similar":

make	+	*effort*	↔	*Anstrengung*	+	*unternehmen*
Auswahl	+	*treffen*	↔	*fare*	+	*selezione*
			↔	*operare*	+	*selezione*
			↔	*effettuare*	+	*selezione*

- Collocations may correspond to single words:

Erfolg	+	*haben*	↔	*succeed*
in Kenntnis	+	*setzen*	↔	*informer*

- Collocates vs. target language syntax:
  - *erinnern* <Subject Object *an*-Object>
  - *in Erinnerung bringen* <Subj. Obj. Ind.Obj.>
  - *rappeler* <Subject Object *à*-Object>

Collocation – Definitional elements from linguistic and lexicographical theory

Towards an integrated approach

<u>Definitional elements</u>

- Two lexemes (N, V, Adj, Adv)
- Syntactic aspect:
  - Noun + Adjective, Noun + Noun
  - Noun + Verb, Noun + Noun
  - Verb + Adverb
  - Adjective + Adverb
- Semantic aspect:
  Collocate is somehow semantically less autonomous than the base ("partial compositionality")
- Pragmatic aspect:
  Collocations "look familiar":
  - Memorized like idioms ?
  - Frequency of the cooccurrence ?
    HAUSMANN: "Halbfertigprodukte der Sprache"

Collocation – Definitional elements from linguistic and lexicography theory

Towards an integrated approach:
The proposal by [Klotz 2000]

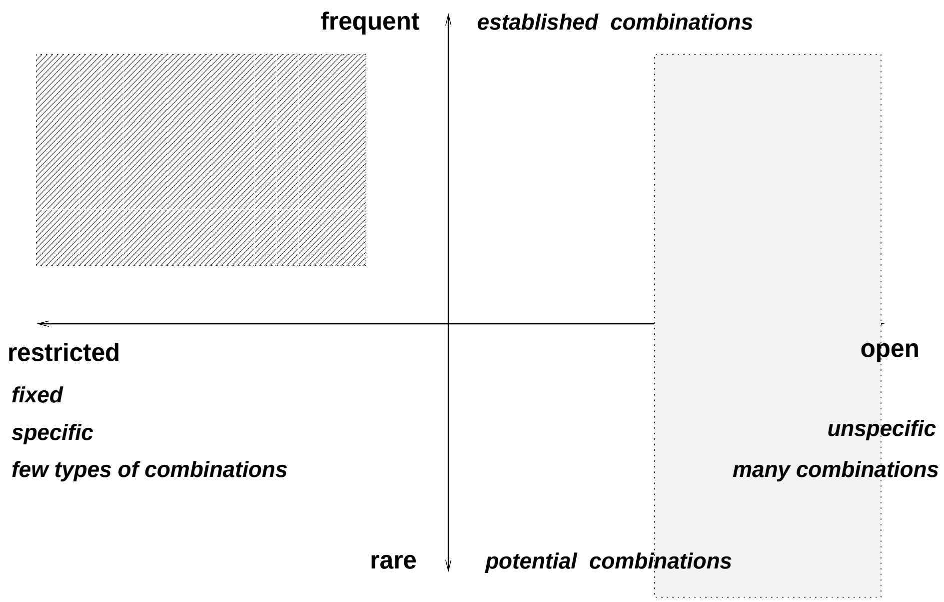
Determining the significance of collocations

Through a combination of both approaches:

- Significance depends on two paramaters:
  - Fixedness ("Restriktivität"):
    - Specific meaning of the collocate
    - Existence of only few combinations (maybe just one)

    ⇒ Idiomaticity
  - Frequency of use ("Etabliertheit"):
    Corpus frequency,
    unless motivated by extralinguistic reasons
- Both paramaters seem to be related:
  Restricted collocations are mostly frequent

Schema



Problems of a lexicographic description of collocations

Collocations vs. other multiword items (1/6)

Lexical vs. grammatical collocation: Collocation vs. Syntax

- Examples from BBI, cf. [Benson 1985]

  proposal *n.*

  1. to make, present, put forth, put forward a ∼
  2. to accept, adopt a ∼
  3. to consider, entertain; receive a ∼
  4. to kill (colloq.), reject, turn down a ∼
  5. a concrete ∼; 6. a ∼ falls through; 7. a ∼ for
  8. a ∼ to + inf. (the committee rejected the ∼ to reduce taxes)
  9. a ∼ that + clause; subj. (they presented a ∼ that all workers be/should be given free dental care)

  propose *v.*

  1. (svo to) she ∼d a new plan to us.
  2. (sv; intr.) to ∼ to ('to offer marriage to')
  3. (sv to INF) I ∼ to leave very early
  4. (sv V-ing) she ∼d leaving very early
  5. (sv poss v-ing) she ∼d his going in my place
  6. (sv that; subj.; to) we ∼d (to them) that she be/should be appointed

- Relationship between syntax and collocation:
  - Collocations (N+V) are usually instances of syntactic subcategorization patterns
  - Need to distinguish between syntactic and collocational properties of words: Benson's grammatical collocations: in fact syntax

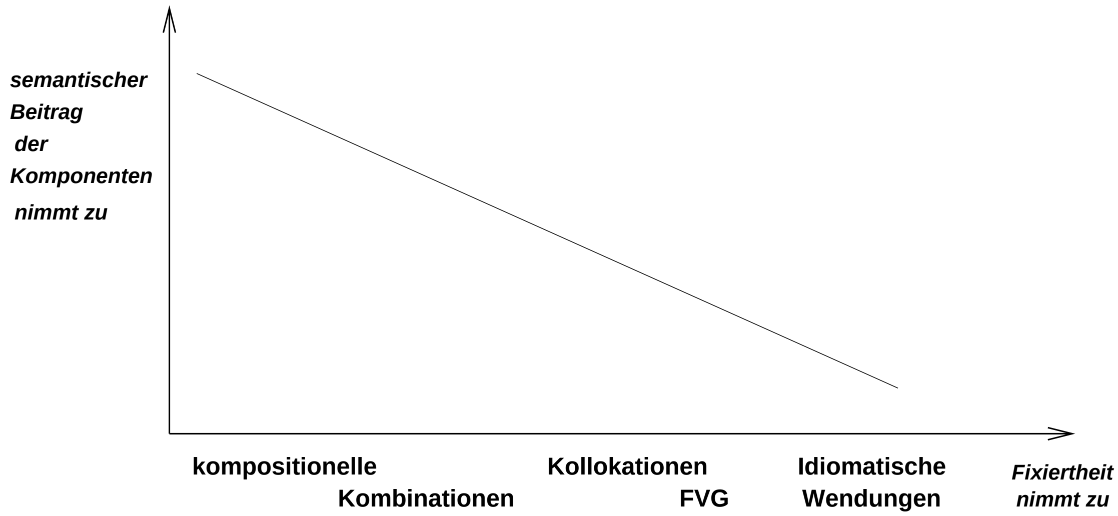
Problems of a lexicographic description of collocations

Collocations vs. other multiword items (2/6)

Collocations on the idiomatic continuum

- Relationship between collocations and
  - Idiomatic expressions (not compositional)
  - Selectional classes
  - Free combinations (fully compositional)
- Specificity (KLOTZ) may come in degrees, and may affect different levels of description:
  - Morphosyntax:
    Restrictions on determination and modification:
    Protokoll schreiben; zur Sprache kommen
  - Semantics:
    aufwerfen in *Frage aufwerfen*

  ⇒ Problem: Borderline with idiomaticity



Problems of a lexicographic description of collocations

Collocations vs. other multiword items (3/6)

Collocations vs. selectional classes (1/2)

- Observation:
  Collocations often do not fit into selectional classes
- Example from verb+object - combinations
  extracted from BNC
  by [Prescher/Riezler/Rooth 2001]:
  [to] cross + Objects (estimated frequencies)

cross +	nouns	f
	mind	74.2
	road	30.3
	line	28.1
	bridge	27.5
	room	20.5
	boarder	17.8
	boundary	16.2
	river	14.6
	street	11.5
	atlantic	9.9

Problems of a lexicographic description of collocations

Collocations vs. other multiword items (4/6)

Collocations vs. selectional classes (2/2)

Towards an automatic classification

- Very often,
  collocations are more frequent
  than the average members of a selectional class
- Proposal
  by [Prescher 2001], [Prescher/Heid 2000]:
  - Extract verb+object pairs
  - Run clustering to get selectional classes
  - Compare frequency with class probability:
    Where difference is highest: collocation candidates
- Examples:
  - Collocation candidates:
    Straftat begehen (`commit a crime`)
    Entscheidung treffen (`make a decision`)
  - "Good" vs. "bad" candidates (verb+subject):
    die Waffen schweigen vs. *Franz spendet*

Problems of a lexicographic description of collocations

Collocations vs. other multiword items (5/6)

Collocations vs. Idioms

"Morphological idiomatization" of noun+adjective groups

- Use in a given case: adverbial NPs

Collocation	Most frequent form	
fern + Tag	eines fernen Tages	(gen)
schön + Tag	eines schönen Tages	(gen)
ganz + Tag	den ganzen Tag	(acc)
jed- + Tag	jeden Tag	(acc)

- Use within PPs: adverbial PPs

Collocation	Most frequent form
Linie + erst	in erster Linie
Linie + ganz	auf der ganzen Linie
Tag + heutig	bis auf den heutigen Tag

- Adjective used in the comparative or superlative

jung + Tag	der jüngsten Tage (gen)
früh + Tag	früherer Tage (gen)
gut + Tag	bessere Tage

⇒ Certain morphological properties seem to contribute to the combination's idiomatic rather than collocational character

Problems of a lexicographic description of collocations

Collocations vs. other multiword items (6/6)

<u>Combinations of collocations vs. idioms</u>

- Combinations of collocations:
  - Adj.+N (*proposal*): *concrete, detailed,..*
  - V+N (*proposal*): *put forward, make,..*
  - Combinations N+Adj.+V:
    - *put forward a concrete proposal*
    - *make a detailed proposal*
- Idiomatic groups N+Adj.+V:
  - *einen (raben)schwarzen Tag haben*
  - *(jdm.) einen guten Tag wünschen*
  - Idiomatic groups with adjectives from a list:

Entwicklung + nehmen:	*gut, schlecht, brillant, ...*
Ausmaße + annehmen:	*riesig, bedrohlich, ...*
Art + haben:	*zupackend, freundlich, ...*
Figur + abgeben | machen:	*traurig, seltsam, gut ...*
in + Stadium + sein:	*entscheidend, früh, letzt, ...*
aus + Verhältnissen + stammen:	*geordnet, einfach, gut,....*

  - ◦ Not decomposable into two collocations
  - ◦ No distinction ‘base vs. collocate’
  - ◦ Sometimes opaque meaning

Lexicographic details about collocations

Types of collocations – by word class

<u>Types relevant for a dictionary</u> (German, English)

<u>Noun</u> + Verb	
— Subject + Verb	*die Frage kommt auf*
— Verb + Object	*Frage + aufwerfen*
— Verb + PP	*zur Sprache kommen*
<u>Noun</u> + Adjective	*stille Hoffnung*
Noun + Noun	
— <u>Noun</u> + Noun$_{Genitive}$	*Tag des Herrn*
— Noun + <u>Noun$_{Genitive}$</u>	*Tor des Tages*
— <u>Noun</u> + Prp + Noun	*knowledge of language*
— Noun + Prp + <u>Noun</u>	*quest for knowledge*
<u>Verb</u> + Adverb	*tief schlafen*
<u>Adjective</u> + Adverb	*tief rot*

<u>Problem: Morphologically related words</u>

- In collocates:
  heavy drinker – heavy drinking – drink heavily
  Stimme abgeben – Abgabe der Stimme – Stimmabgabe
  ⇒ How many collocations?
- In bases:
  Pause einlegen – Atempause einlegen
  ⇒ Cf. talk by Heike Zinsmeister

Lexicographic details about collocations

Examples of relevant information types (1/3)

Objective

In a collocation dictionary
or in a detailed monolingual dictionary for text production,
collocations should be described
in the same detail as single word lemmas.

Descriptive programme (N+V-Collocations)

♯	Obj.	Level	Property	Example
1	Col.	lexical	lexeme cooccurrence	*Rede + halten*
2		syntactic	subcat. of collocation	*in der Lage sein zu + INF*
3		semantic	aktionsart, aspect	"INCEP OPER$_1$ (*Ansicht*)"
4			synonymy	*ins Schwitzen kommen | geraten*
5		pragmatic	diasystematic marks	*Sorge tragen zu* (swiss)
6			frequency	f (*Vorschlag machen*) >> f (*Vorschlag unterbreiten*)
7	N.	morphosynt.	number	(jmdm.) *Grenzen setzen* (plur.)
8		syntactic	determination	*zu der Ansicht gelangen, dass...* definite, non-fused article (**zur Ansicht gelangen, dass*)
9			modification potential	*eine ? [ ] | gute Entwicklung nehmen*
10			subcategorization	*zu der Ansicht kommen, dass...*
11	V.	morphosyntactic	active/passive	*Eile ist geboten* (passive)
12			tense preferences	
13		syntactic	subcategorization	*zu der Ansicht kommen*

Lexicographic details about collocations

Examples of relevant information types (2/3)

Number of the noun in N+V-collocations

- The case of *Hoffnung(en)*

```
noun form    verb lemma     Freq./Cp.  LogLikelihood
-------------------------------------------------------
Hoffnungen   sich_machen     74 HGC       724.19
Hoffnung     sich_machen     21 HGC       107.69

Hoffnung     trügt           37 HGC       347.60
Hoffnungen   trügen           5 HGC        34.37

Hoffnung     schöpfen        41 HGC       507.56
```

- The case of *Grenze(n)*

Collocation	more sg	more pl
Grenzen abschotten		+
Grenzen dichtmachen		+
Grenzen schließen		+
Grenzen aufzeigen		+
Grenzen ausloten		+
Grenze überqueren	+	
Grenze passieren	+	

Lexicographic details about collocations

Examples of relevant information types (3/3)

The determiner of the noun in N+V-collocations

- Fixed determination (e.g. in Support Verb Constructions)
  - Determiners (definite, indefinite, null)
    often fixed, as part of the idiomatization:
    { } *Ausdruck finden*, { } *Protokoll schreiben*
  - Article + Preposition: fused vs non-fused:
    zum Erliegen kommen,
    zu dem Schluß kommen, daß ...

non-fused zu + ART	fused *zum, zur*	both
	Anwendung	
	Aufführung	
	Besinnung	
Einigung		
	Entfaltung	
Entscheidung		
Erkenntnis		
		Explosion
	Geltung	
		Katastrophe

- Preferences, e.g. for possessives:
  - *seine Vorstellungen einbringen*
  - *sein Veto einlegen*
  - *sein Ende finden*

Collocations in dictionaries

Linguistic classification vs. user needs

Elements of the linguistic classification

Reminder: Specificity of collocations vs. Idiomaticity

- Collocation vs. Element of a selectional class
- Collocation vs. two-word idiom:
  Morphological idiomatization
- Combination of collocations vs. idiom

Questions

- Which impact has the linguistic classification
  on the description and presentation
  of multiword items in dictionaries?
- Can dictionary users distinguish
  between collocations and idioms?
- Do users need to distinguish?

Current situation in dictionaries

- Most dictionaries
  have no specific item type for collocations
- Example syntagms in monolingual dictionaries
  are polyfunctional:
  They illustrate syntax and collocation(s) at a time

Collocations in dictionaries

User needs – Dictionary functions –
and the presentation of collocations

Basic types of usage situations
Based on [Bergenholtz 2002]

- Problems of text understanding
- Problems of text production
- Problems of text translation
- Need for knowledge about words

More detailed typology (text-related dictionary use)

Functions of dictionary	Lang. of user	Lang. of text
Text understanding	L_1	L_1
Text production	L_1	L_2
Text translation	L_2	L_1
	L_2	L_2

$\Rightarrow$ A total of 12 constellations

Collocation-related needs
Example case: L_1 user produces texts in L_2:
– Collocates must be listed under Bases
– Collocations must be distinguished from free combinations
– Collocations must be distinguished from idioms

Collocations in dictionaries

The current situation: Monolingual dictionaries

Polyfunctional items

- *Duden Universalwörterbuch* ^{4}*2001*:
  - Free combinations and collocations: same layout and font
  - Examples contain (combinations of) collocations
  - **Vorschlag** [...] jmdm einen V. unterbreiten; praktische Vorschläge machen; einen V. annehmen, akzeptieren, aufgreifen, billigen, erwägen, prüfen, ablehnen, befolgen; ich mache dir einen V. (*ich schlage dir Folgendes vor*); sie ging auf seinen V. nicht ein; über einen V. abstimmen.
- *Duden. GWDS* ^{2}*1995*:
  Citations contain sometimes collocations,
  sometimes special uses created by the author

  Vorschlag [...]

  Deshalb erlaube sie sich ... folgenden ... V. zu erstatten (Musil, Mann 179); er bereute tief, Giovannis wahnwitzigen V. befolgt zu haben (Thieß, Legende 27)

Needs

- Distinction should be made
  between collocational uses reusable in text production
  and other uses, relevant e.g. for understanding
- Collocations should not be combined

Cf. now: [Köster/Neubauer 2002] on Kempcke

Collocations in dictionaries

The current situation: Bilingual dictionaries

No specific collocation items

- Collocations are found in examples, as well as in translation-related glosses
- In addition: Problems of polyfunctionality of examples, as in monolingual dictionaries
- Glosses are poor in syntactic information, not always useful for text production

Examples

Langenscheidts Handwörterbuch Deutsch/Englisch

- **berufen I.** v/t **1. j-n zu** *einem Amt* ~ appoint s-o to; **j-n zum Vorsitzenden** ~ appoint s-o chairman
- **nähren I.** v/t **1.** feed; **2.** *fig. (Hoffnung etc.)* nurture, *(Haß, Verdacht, etc.)* a. fuel

Collocations in dictionaries

Is the linguistic classification useful? (1/2)

Collocations vs. selectional classes

- Text production:
  Distinction is necessary to allow for extrapolation:
  cross o's mind – cross the street/the ocean/...
  Indication "... etc." is often sufficient
- Text understanding:
  Distinction is not really needed

Collocations vs. two-word idioms

- Text production:
  Necessary for correct use;
  Often derivable from word class indication
- Text understanding:
  Distinction is not really needed

Combinations of collocations

- Text production:
  Examples with combined collocations are dangerous:
  is the fact meaningful
  that a combination of collocations is given?
- Text understanding:
  Combinations make collocations more difficult to find

Collocations in dictionaries
Is the linguistic classification useful? (2/2)

Overall

- The classification allows to distinguish specific from more open types of collocations
- Classification has an impact on the presentation of collocational data
- This in turn facilitates the interpretation of collocational indications in a given dictionary

Corpus lexicographic tasks concerning collocations

The case of dictionary updating – TFB 32

Objectives

- Proposals for additions to and (maybe) removals from collocational information of a given dictionary
- Collecting material on collocation candidates:
  - Lexical cooccurrence, example sentences
  - Frequency data and data from Association Measures
  - As much classificatory data as possible
- Presentation as if it were intended
  for text production and text understanding
  ⇒ List collocations twice

Procedures (Overview)

- Extraction by means of two tools:
  - Chunk-based patterns: N+Adj, Adj+Adv, N+N
  - Stochastic grammar (Gramotron): N+V, V+Adv

  ⇒ Provides grammatically homogeneous material
  ⇒ Provides classificationally relevant features
- Data are sorted, by bases and collocates, resp.

Corpus lexicographic tasks concerning collocations

The case of dictionary updating – Open questions

Problems

- For some words, there are no significant collocates
- Otherwise, lists of candidates are very long:
  Where to cut them off?

Corpus lexicographic tasks concerning collocations

Examples of results from Gramotron

Lists by cumulative frequency

- Base: *Hoffnung*

180.17	machen	13.58	enttäuschen
175.36	setzen	13.31	verlieren
155.2	haben	13.00	hegen
116.82	geben	10.98	begraben
72.74	aufgeben	10.93	nehmen
55.51	schöpfen	10.20	erfüllen
35.98	dämpfen	10.01	nähren
31.95	wecken	9.42	zerstören
24.99	äußern	9.00	teilen
16.81	verbinden	5.3	bringen

- Collocate: *hegen* (plus Object nouns)

53.85	Zweifel	3.00	Interesse
17.99	Plan	3.00	Illusion
16.00	Verdacht	2.99	Vorstellung
13.99	Befürchtung	2.00	Ressentiment
13.00	Hoffnung	2.00	Utopie
11.32	Gefühl	1.00	Zuneigung
10.92	Ambition	1.00	Respekt
7.96	Sympathie	1.00	Begeisterung
5.57	Bedenken	1.00	Haß
5.09	Erwartung	1.00	Furcht
5.00	Mißtrauen	1.00	Absicht
5.00	Traum	1.00	Bewunderung
4.99	Wunsch	1.00	Aversion
4.00	Gedanke	1.00	Verständnis
4.00	Groll	1.00	Mitleid
		1.00	Idee

Corpus lexicographic tasks concerning collocations

Examples of results from dictionary updating:
Comparing corpus and dictionary

Examples from an upcoming bilingual dictionary

Entry	Adjective	Noun	in dict	Cp
ungebärdig	ungebärdiges	Pferd	yes	0
ungebärdig	ungebärdiger	Gebirgsfluß	yes	0
ungebärdig	ungebärdiger	Sprößling	no	5
frei	freier	Mitarbeiter	yes	53
frei	freies	Feld	yes	50
frei	freier	Wille	yes	12
	freier	Markt	no	139
	freie	Marktwirtschaft	no	50
	freie	Fahrt	no	138
	freie	Meinungsäußerung	no	59
	freie	Hand	no	62
	freien	Lauf (lassen)	no	97

dict.:	Licht	abblenden	from Corp.:	Licht	abstrahlen
dict.:	Licht	abdunkeln	from Corp.:	Licht	anknipsen
dict.:	Licht	abschalten	from Corp.:	Licht	anmachen
			dict.:	Licht	anstellen

⇒ Some examples from the dictionaries are not found in the corpus

⇒ Some collocation candidates frequent in the corpus are not (yet) in the dictionary

Corpus lexicographic tasks on collocations

Problems with the use of candidate lists
Items without collocations?

The case of *erforschen* and its objects

- *Duden 2. Stilwörterbuch*:

  erforschen [...] den Weltraum e.; das verhalten von Tieren e.; den Menschen e.; die Antarktis [...]; die Hintergründe e.; sein Gewissen e. (*prüfen*); die Wahrheit über etw. e.

- Data from Gramotron:

```
erforschen
        Frequenz:
        Meinung           5.00
        Geschichte        4.00
        Entwicklung       4.00
        Auswirkung        3.00
        Land              3.00
        Hintergrund       2.99
        Möglichkeit       2.81
        Reaktion          2.81
        Leben             2.62
        Historie          2.00
        Phänomen          2.00
        Vergangenheit     2.00
        Wechselwirkung    2.00
        Insel             2.00
        Welt              2.00
        Klarheit          2.00
        Vorgang           2.00
        Verseuchung       2.00
        Kultur            2.00
        Teil              1.99
        System            1.99
        Limes             1.99
        Prozeß            1.99
        Stimmung          1.93
        Bedingung         1.91
        Ökosystem         1.84
        Wirkung           1.82
        In                1.60
        Fähigkeit         1.00
        Brauchbarkeit     1.00
        Pfad              1.00
```

Possible interpretation

- Any noun lemma is possible as an object
- No single object noun is particularly probable
- Lexicographic description:
  These are not collocations, but trivial combinations

Corpus lexicographic tasks on collocations

Problems with the use of candidate lists

Heuristics for delimiting significant collocations

Candidate lists (s.v. *Vorschlag*)

```
Vorschlag
        t-score:
        machen              15.59
        unterbreiten        13.63
        ablehnen            13.19
        vorlegen            10.27
        erarbeiten           7.84
        prüfen               7.83
        begrüßen             7.29
        unterstützen         6.41
        zurückweisen         6.39
        einbringen           6.18
        ausarbeiten          6.11
        aufgreifen           6.05
        diskutieren          5.68
        akzeptieren          5.56
        präsentieren         5.27
        bezeichnen           5.01
        billigen             4.61
        kritisieren          4.36
        enthalten            4.28
        annehmen             3.90
        stoßen               3.90
        aufnehmen            3.71
        berücksichtigen      3.52
        umsetzen             3.45
        einreichen           3.37
        beraten              3.25
        entwickeln           3.18
        halten               3.01
        verwirklichen        2.93
        begründen            2.67
        eignen               2.65
        besprechen           2.63
        zurückziehen         2.60
        abstimmen            2.59
    Frequenz:
        machen             353.98
        ablehnen           191.80
        unterbreiten       187.14
        vorlegen           114.63
        prüfen              67.12
```

Proposals (Heike Zinsmeister, TFB 32)

- Frequency-based cut-off
- Use of t-score or LogLikelihood for sorting

Corpus lexicographic tasks on collocations

Problems with the use of candidate lists
Heuristics for delimiting significant collocations

Proposals for sorting the candidates

- Frequency-based cutoff: e.g. $f > 5$
- Use of Association Measures,
  e.g. t-score or LogLikelihood, for sorting the candidates
- Threshold t-score: e.g. $t > 2.57$
- Practical proposal:
  - Offer lexicographers a short list of "good" candidates
  - Allow access to other data, e.g. on a frequency basis

Summary

This presentation
A collection of
a few relevant points about collocations and lexicography

- Two lines of tradition in lexicography and linguistics:
  — Both have influenced corpus based work and NLP
  — A combined approach seems feasible
- Descriptive and presentational tasks for lexicography:
  — Collocations need a detailed description
  — Separation from idioms is necessary
- Corpus lexicographic work in TFB 32:
  Open problems:
  Concerning the delimitation from trivial items