EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT
FACHBEREICH INFORMATIK

Lexikologie: Grundbegriffe

Version 2019-10-18

Timm Lichte

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Outline

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Intro I

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lexikologie

Die Lexikologie innerhalb der Linguistik beschäftigt sich mit dem Bestand der Worte einer Sprache und zielt darauf ab, deren Eigenschaften (insb. Bedeutung) und Beziehungen untereinander zu erfassen.

- Tatsächlich geht es in der Lexikologie nicht nur um “Worte” im engeren Sinn (siehe unten), sondern auch um größere lexikalische Einheiten wie ***ins Gras beißen***.

Benachbarte Bereiche (andere Folien):

- Die Struktur der Worte wird in der **MORPHOLOGIE** untersucht.
- Die Verknüpfung der Worte wird in der **SYNTAX** untersucht.
- Die Bedeutung von Worten, Sätzen und Diskursen wird in der **SEMANTIK** untersucht.

Intro II

In diesem Foliensatz werde ich ein paar grundlegende Begriffe (englische Namen in Klammern) der Lexikologie definieren:

- Token (token)
- Typ (type)
- Wortform (word form)
- Basis- und Zitationsform (base and citation form),
- Lemma (lemma)
- Lexem (lexeme)
- Lexikon/Wörterbuch (lexicon/dictionary)
- Sprache (language)

Die Definitionen sind allesamt “extensional”, d. h. die Definitionen stellen eine explizite Verbindung zu einem Modell mit einer Menge mehr oder weniger konkreter Objekte, her.

Allgemeines Wortverständnis I

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Die Definitionen setzen allesamt ein enges Wortverständnis voraus:

Worte (allgemein)

Worte sind die minimalen Zeichenketten, die in wohlgeformten Sätzen durch Leerzeichen abgrenzbar sind und selber keine Leerzeichen enthalten.

- ***Gras*** ist ein Wort, ***ins Gras beißen*** nicht.
- ***ins*** ist ein Wort, ***s*** aber nicht (obwohl es ein Überbleibsel des Worts ***das*** ist).
- ***Spiegelei*** ist ein Wort, ***Spiegel-*** als erste Komponente des Kompositums ***Spiegelei*** aber nicht.
- Interpunktion ist dagegen kein Wort, da Punkt, Komma, etc. in einem (korrekt geschriebenen Satz) **nie** in dieser Weise abgrenzbar sind.

Allgemeines Wortverständnis II

Der Begriff “Wort” wird gemeinhin aber auch so verwendet:

- “Wieviele **Worte** hat der Text?”
- “Ich würde hier ein anderes **Wort** benutzen.”
- “Bei diesem **Wort** hast du dich verschrieben.”
- “Konntest du das **Wort** im Lexikon finden?”

In diesen Sätzen kann “Wort” unterschiedliche Bedeutungen haben:

- ***Wordform*** (≈ Abfolge von Buchstaben)
- **`Wordtoken`** (≈ Wortform mit Position)
- LEMMA (≈ Menge von Wortformen)
- LEXEM_1 (≈ Lemma + Bedeutung)

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Token versus Typ I

- Die Begriffe Token und Typ sind immer **bezogen auf ein Modell** zu verstehen, d.h. auf eine gegebene Menge von Objekten, Objektmengen und Objektrelationen.
- Objekte eines Modells können etwa die Buchstaben(-Instanzen) auf dieser Folie sein, wobei jeder Buchstaben nicht nur anhand der **Form**, sondern auch anhand der **Position** von anderen Buchstaben unterschieden wird. Diese Objekte nennt man auch **TOKEN**.
  - Beispiel: `Bierdeckel` enthält 10 Objekte/Token, nämlich B_1, i_2, r_3, usw.
- Der Teufel steckt aber wie immer im Detail: Was **genau** ist unser Modell? Wie konkret oder abstrakt soll es sein? Wir werden gleich bei der Definition von **WORTTOKEN** auf diese Frage zurückkommen.

Token versus Typ II

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Andererseits kann man Objekte/Token zusammenfassen zu Mengen. In obigen Modell könnten wir z.B. Mengen bilden, die abstrakten Buchstaben entsprechen. Solche Mengen nennen wir **TYPEN**.
  - Beispiel: `Bierdeckel` enthält 8 Buchstabentypen: {B_1}, {i_2}, {e_3, e_6, e_9}, {r_4}, {d_5}, {c_7}, {k_8}, {l_{10}}

Token versus Typ III

Etwas formaler ausgedrückt können wir sagen: Gegeben ein Modell $\mathcal{M}$ mit einer Domäne $\Delta^{\mathcal{M}}$.

Token

Ein Token bezogen auf $\mathcal{M}$ ist ein Objekt in $\Delta^{\mathcal{M}}$.

Typ

Ein Typ bezogen auf $\mathcal{M}$ ist eine Teilmenge von $\Delta^{\mathcal{M}}$.

Wenn $\Delta^{\mathcal{M}} = \{x_1, ..., x_n\}$ also die Menge der Buchstabentoken auf dieser Seite ist, dann könnte z.B. $\{x_5, x_{23}, ...\} \subset \Delta^{\mathcal{M}}$ der Buchstabentyp sein, der allgemein den Buchstaben ***s*** repräsentiert.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Worttoken I

Entsprechend können wir die Token/Typ-Unterscheidung auch bei Modellen anwenden, in denen die Objekte Worte sind.

Worttoken (word token)

Ein Worttoken ist ein konkretes, sinnlich wahrnehmbares Wort mit raumzeitlichen Koordinaten.

- Das Wort **`Bierdeckel`** , das gerade auf Ihrem Monitor dargestellt wird, oder das Sie auf Ihrem Ausdruck lesen, ist ein Wordtoken. Dagegen ist **`Bierdeckel`** ein anderes Wordtoken: es steht anderer Stelle als das erste **`Bierdeckel`**, womit Sie dann schon drei von der Form her identische Wordtoken vor sich haben . . .
- Das Spiel kann man aber noch weiter treiben: bei **`Bierdeckel`** kann man selbst dann unendlich viele Worte sehen, wenn es an einem Ort erscheint, und zwar wenn man zusätzlich die Worte auf einer realwertigen Zeitachse verortet.

Worttoken II

- So weit möchte man in der Regel nicht gehen. Man einigt sich (mehr oder weniger bewußt) auf eine Wordtokendomäne bzw. ein Modell, das von solchen realwertigen physikalischen Dimensionen abstrahiert.

Normalerweise nimmt man an, dass ein Modell für diesen Foliensatz auf die **Art** seiner digitalen Representation Bezug nimmt, die (grosso modo) auf jedem Rechner gleich sein sollte und zeitunabhängig ist. Wir bezeichnen Modelle mit dieser Eigenschaft im Folgendem als $\mathcal{M}^D$ (gewissermaßen ein Modell-**Typ**).

Worttokenmodell $\mathcal{M}^w$

$\mathcal{M}^w$ ist das Worttokenmodell bezogen auf die digitale Representation einer Sprachressource/Korpus $\mathcal{M}^D$.

Worttoken III

- Beispiel: Nehmen wir an, dass die Sprachressource aus dem Satz ***Peter knickte den Bierdeckel*** besteht, dann ist das Wordtokenmodell $\mathcal{M}^w = \{$ ***Peter***$_1$, ***knickte***$_2$, ***den***$_3$, ***Bierdeckel***$_4$ $\}$.

Wenn wir schon so ein Wortokenmodell $\mathcal{M}^w$ haben, müssen wir keinen Unterschied mehr zur Wordtokendomäne $\Delta^{\mathcal{M}^w}$ machen:

Worttokendomäne $\Delta^{\mathcal{M}^w}$

$\Delta^{\mathcal{M}^w}$ ist die Worttokendomäne von $\mathcal{M}^w$. Es soll gelten: $\Delta^{\mathcal{M}^w} = \mathcal{M}^w$.

So verstanden, sind Wordtoken ein wichtiges Maß bei der Bestimmung der Größe (aber nicht der Entropie) von Textkorpora (d.h. der Art ihres digitalen Gegebenseins . . .)

- Ein Textkorpus kann aus *n* Worttoken bestehen, die alle die Form ***Bierdeckel*** haben, obwohl das nicht sehr realistisch ist.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Worttyp & Worttypendomäne

Entsprechend der Token/Typ-Unterscheidung oben und unter Rückgriff auf das Modell $\mathcal{M}^w$ eines Textes auf Grundlage seiner digitalen Repräsentation, können wir definieren:

Worttyp

Ein Worttyp in einem Modell $\mathcal{M}^w$ ist eine Menge von Wordtoken aus $\Delta^{\mathcal{M}^w}$.

D.h. die Menge der möglichen Worttypen ist die Potenzmenge der Menge der Wordtoken, also $\mathcal{P}(\Delta^{\mathcal{M}^w})$. Wir bezeichnen die Menge von Worttypen, die für $\mathcal{M}^w$ angenommen werden, als **WORTTYPENDOMÄNE** von $\mathcal{M}^w$.

- Beispiel: Nehmen wir an, dass $\mathcal{M}^w$ aus 3 Wordtoken der Form ***Bierdeckel*** besteht. Dann gibt es dazu ein Worttyp {**Bierdeckel**$_1$, **Bierdeckel**$_2$, **Bierdeckel**$_3$ }.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Wortformen I

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Wir haben jetzt alles, um zwei wichtige Worttypendomänen zu definieren: Wortformen und Wortbedeutungen.
Gegeben ein Modell $\mathcal{M}^w$ mit einer Worttokendomäne $\Delta^{\mathcal{M}^w}$:

Wortformen $\hat{\Delta}_{\phi}^{\mathcal{M}^w}$

Die Menge der Wortformen $\hat{\Delta}_{\phi}^{\mathcal{M}^w}$ ist die kleinste Menge derjenigen Worttypen in $\mathcal{M}^w$, deren Wordtoken dieselbe Form haben.

“Form” wird intuitiv verstanden als Konfiguration von Buchstabentypen. Also: `Bierdeckel` und `Bierdeckel` haben diesselbe Form , aber nicht `Bierdeckel` und `Bierdeckels` (die Genitivform).

Wir schreiben: `Bierdeckel` und `Bierdeckel` haben diesselbe Wortform ***Bierdeckel***.

Wortformen II

Die Definition einer Wortform sieht dann so aus:

Wortform (word form)

Eine Wortform bezogen auf $\mathcal{M}^w$ ist ein Element der Worttypendomäne $\hat{\Delta}_{\phi}^{\mathcal{M}^w}$.

Wortbedeutungen I

Wir brauchen den Begriff der Bedeutung gleich bei der Definition von Lexemen.

Der zentrale Begriff “Bedeutung” ist natürlich sehr schwierig zu bestimmen (anders als die Form eines Wortes). Ich werde hier die Intuition des **KONZEPTUELLEN GEHALTS** zugrunde legen, mit dem sich die Bedeutung zweier Worttoken zumindest **vergleichen** lässt. Beispiel: `Bierdeckel` und `Bier` bedeuten **wahrscheinlich** (das sind ja nur zwei alleinstehende Worttoken) konzeptuell etwas sehr unterschiedliches; dagegen **können** `Bierdeckel` und `Deckel` etwas sehr ähnliches bedeuten.

⇒ Man sieht: Es ist schwierig über Bedeutung allgemein etwas zu sagen, ohne einen bestimmten, mehr oder weniger konkreten Verwendungskontext mitzudenken.

Wortbedeutungen II

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Eine wichtige Frage der Lexikologie, die wir hier nicht allgemeingültig beantworten können, ist:

Wann sind die Bedeutungen so unterschiedlich, dass man von unterschiedlichen Bedeutungen sprechen kann/muss?

Wortbedeutungen III

Um im Folgenden etwas in der Hand zu haben, werden wir Wortbedeutungen als Worttypen definieren.

Wortbedeutungen $\hat{\Delta}_{\sigma}^{\mathcal{M}^w}$

Die Menge der Wortbedeutungen $\hat{\Delta}_{\sigma}^{\mathcal{M}^w}$ ist die kleinste Menge derjenigen Worttypen in $\mathcal{M}^w$, deren Worttoken dieselbe Bedeutung haben (also synonym sind).

Es wird sich im Folgenden als praktisch erweisen, Wortbedeutung gleich als Menge von Wortformen statt als Menge von Worttoken zu verstehen.

- Beispiel: Die Wortbedeutung ‘Finanzinstitut’ ist eine Menge von Wortformen $\{$ ***Bank***, ***Banken***, ***Finanzinstitut***, . . . $\}$, die **irgendwo** in $\mathcal{M}^w$ die konzeptuelle Bedeutung von ‘Finanzinstitut’ haben.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Lemma (lemma) & Zitationsform (citation form) I

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lemma

Eine Lemma ist eine Teilmenge von $\hat{\Delta}_{\phi}^{\mathcal{M}^w}$, d.h. eine Menge von Wortformen.

Als **Bezeichner/Label** für ein Lemma dient eine Wortform des Lemmas. Wir nennen diesen Bezeichner **ZITATIONSFORM (CITATION FORM)**, aber üblich ist auch **BASISFORM (BASE FORM)** oder einfach **LEMMA (LEMMA)**. D.h. ohne Lemma keine Zitationsform!

- Beispiel: Die Wortformenmenge {***Bierdeckel***, ***Bierdeckels***, ***Bierdeckeln*** } bildet ein Lemma mit der Bezeichnung BIERDECKEL.

Der Bezeichner eines Lemmas ist meist die kürzeste Wortform im Lemma (deshalb “Basisform”), aber das unterliegt letztlich der Konvention.

Lemma (lemma) & Zitationsform (citation form) II

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

- Im Deutschen ist die Zitationsform bei Verben der Infinitiv: “Das Verb Gehen . . . ”

Was gehört in ein Lemma? Intuitiv ist das klar, aber wie können wir das genau festlegen? Das ist gar nicht so einfach. Was wir grob sagen können:

- In einem Lemma sind Wortformen mit derselben Bedeutung abgesehen von: Numerus, Geschlecht, Tempus, . . .
- In einem Lemma sind daher Wortformen genau einer Wortart (Nomen, Verb, Adjektiv, . . .).
- In einem Lemma sind Wortformen, die denselben **STAMM** haben. Oben war das die Wortform ***Bierdeckel***.

Was genau ein Stamm ist und wie Lemmata genau gebildet werden, untersucht u.a. die **MORPHOLOGIE**.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Lexem (lexeme) I

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Ein Lexem ist schließlich eine Kombination eines Lemmas mit einer Wortbedeutung. Da wir letztlich alles auf der Grundlage eines Models $\mathcal{M}^w$ mit derselben Tokendomäne definiert haben, können wir die Definition elegant mengentheoretisch aufschreiben:

Lexem (mengentheoretisch)

Ein Lexem L_i ist der nichtleere Schnitt der Wortformen eines Lemmas L mit den Wortformen einer Wortbedeutung σ, d. h. bezogen auf $\mathcal{M}^w$ gilt: $L_i = L \cap \sigma$ mit $L \subseteq \hat{\Delta}_{\phi}^{\mathcal{M}^w}$ und $\sigma \in \hat{\Delta}_{\sigma}^{\mathcal{M}^w}$.

Man kann das griffiger als Tupel aufschreiben:

Lexem (Tupelschreibweise)

Ein Lexem L_n ist das Tupel $\langle L, \sigma \rangle$ bestehend aus einem Lemma L und einer Wortbedeutung σ.

Lexem (lexeme) II

Es gilt die Konvention, dass der Bezeichner für das Lexem abgeleitet ist aus dem Bezeichner des Lemmas.

- Beispiel: Das Lexem BANK$_1$ besteht aus dem Lemma BANK = {***Bank***, ***Bänke***, ***Bänken***, ***Banken***} und der Bedeutung σ_{23} = 'Finanzinstitut'. Dagegen besteht das Lemma BANK$_2$ aus dem Lemma BANK = {***Bank***, ***Bänke***, ***Bänken***, ***Banken***} und der Bedeutung σ_{75} = 'Sandbank'.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Lexikon (lexicon) & Sprache (language) I

Schließlich noch zwei Begriffe, die sich trivialerweise aus dem bisher gesagten ableiten lassen: Lexikon und Sprache.

Lexikon (lexicon)

Ein Lexikon L für $\mathcal{M}^w$ ist eine Menge von Lexemen bezogen auf $\mathcal{M}^w$, so dass jedes Wordtoken in $\mathcal{M}^w$ in der Extension mindestens eines Lexems in L ist.

Ich habe mit Bedacht nicht von **dem** Lexikon und **den** Lexemen gesprochen, denn das lässt das Modell nicht zu. Das liegt an den Wortbedeutungen, die hier frei gewählt werden können (im Sinne ihrer Extension). Die Lemmata scheinen dagegen *relativ* genau festlegbar zu sein, weil die sich aus der Wortform ergeben.

Lexikon (lexicon) & Sprache (language) II

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Über Sprache haben wir eigentlich die ganze Zeit schon gesprochen:

Sprache (language)

Eine Sprache ist ein Modell $\mathcal{M}^w$.

Eine direkte Folge aus den Definitionen von Lexikon und Sprache ist: Für jede Sprache gibt es mehrere mögliche Lexika, wenn man die Wortbedeutungen nicht festlegt. Eine **triviale Festlegung** wäre: Jede Wortbedeutung enthält genau ein Worttoken.

Lexikon (lexicon) & Sprache (language) III

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Wenn wir **Sprachdaten** betrachten, dann gibt es immer mehr als ein mögliches Lexikon, denn die Wortbedeutungen sind nicht vorgegeben.

Wenn wir eine **mentale Repräsentation** der Sprachdaten hinzunehmen, dann gibt es in jedem Kopf möglicherweise nur ein Lexikon. Allerdings ist das ein sehr “lokales” Lexikon.

Man kann dann von “dem” Lexikon einer Sprache nur sinnvollerweise sprechen, wenn man eine extensionale Identität aller “lokalen” Lexika annimmt. Das ist eine nützliche und vielleicht unvermeidbare Abstraktion – realistisch ist das aber nicht.

Outline

1 Intro

2 Token versus Typ

3 Worttoken

4 Worttyp & Worttypendomäne

5 Wortformen & Wortbedeutungen

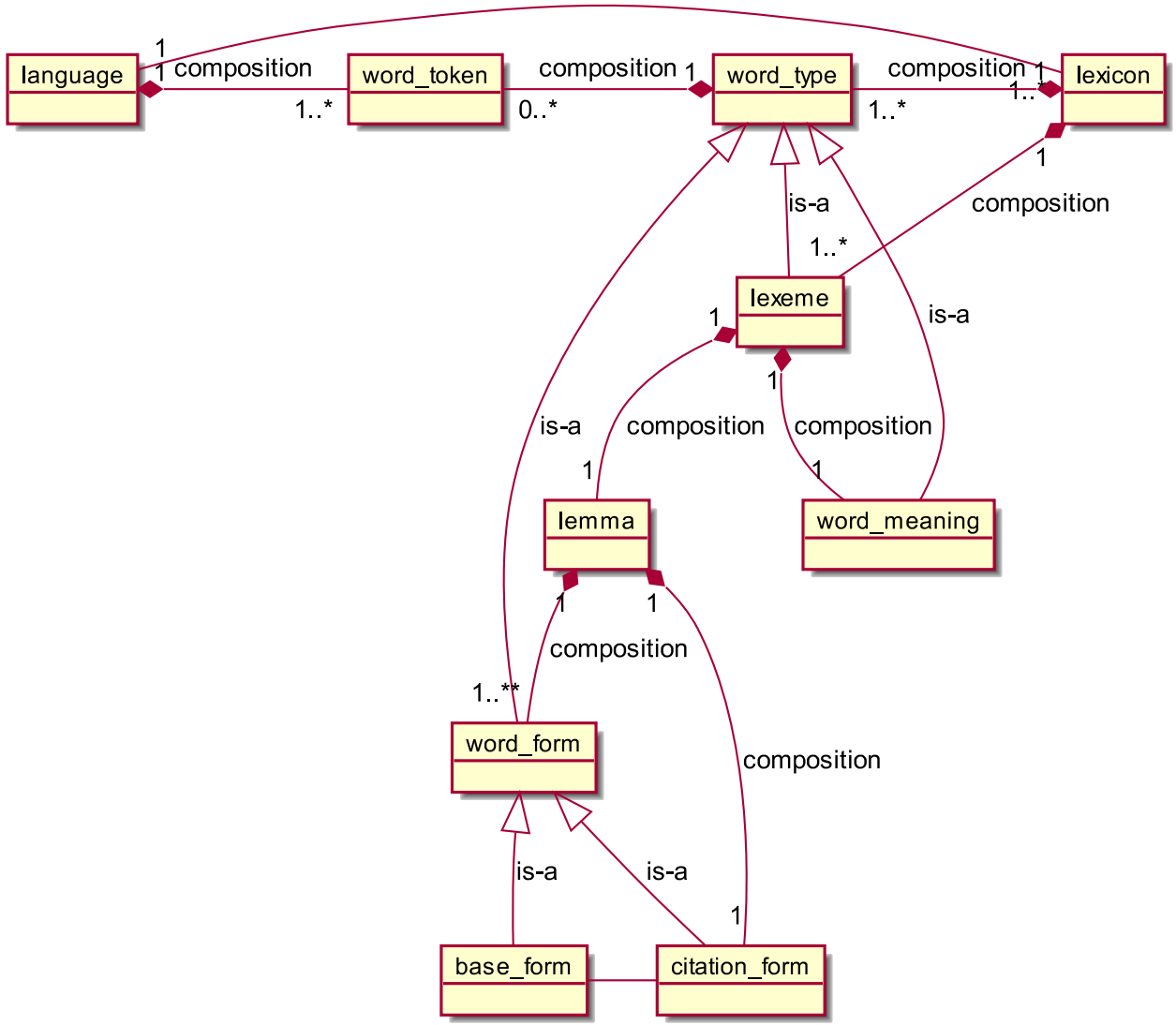
6 Lemma (lemma) & Zitationsform (citation form)

7 Lexem (lexeme)

8 Lexikon (lexicon) & Sprache (language)

9 Begriffsdiagramm

Begriffsdiagramm



Literaturangaben I