

Einführung

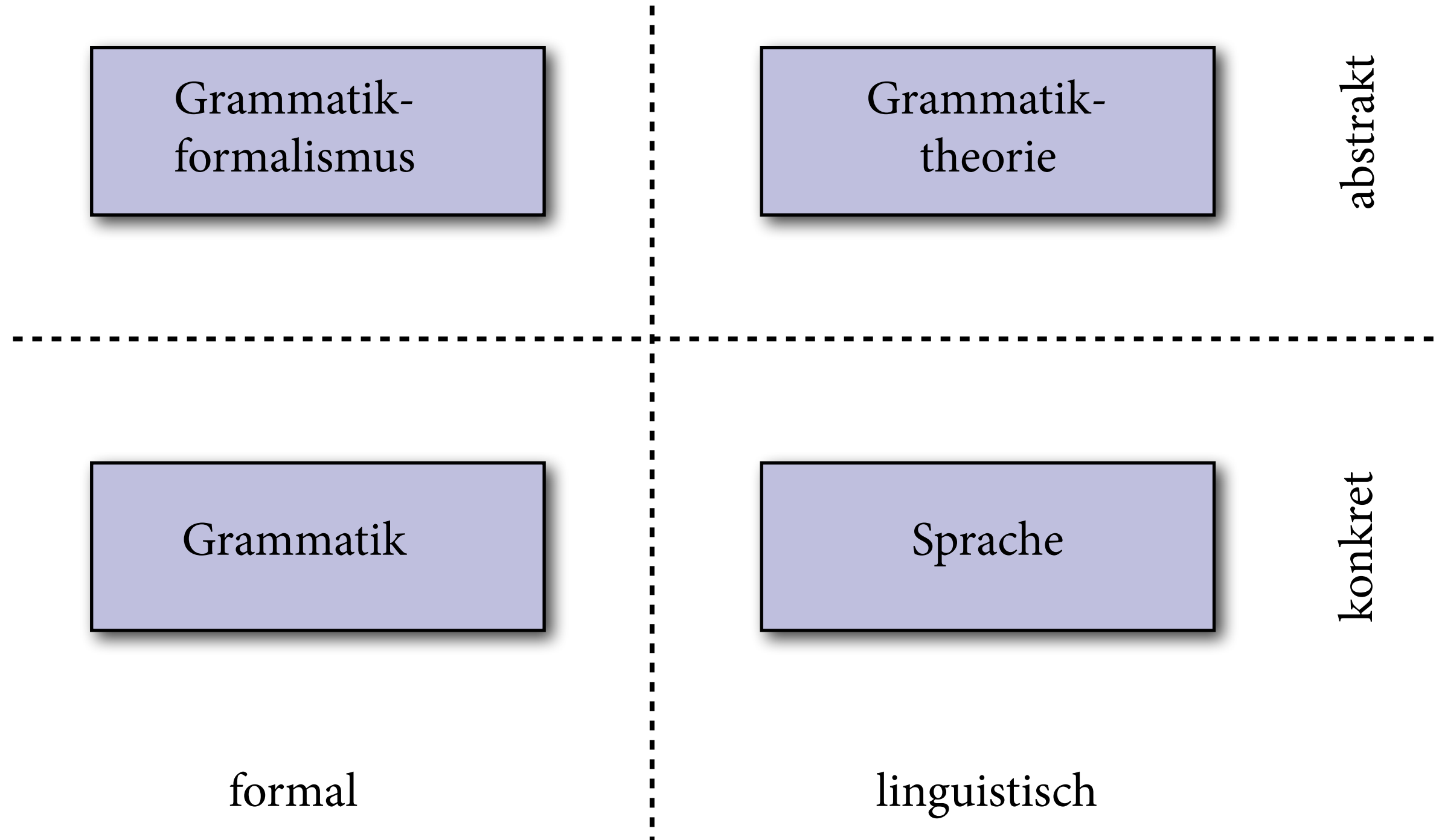
Vorlesung “Grammatikformalismen”
Alexander Koller

12. April 2019

Übersicht

- Was sind Grammatikformalismen, und warum reichen uns kfGs nicht?
- Was machen wir in dieser Vorlesung?
- Prüfungsmodalitäten

Grammatikformalismen



Ein Beispiel

$T = \{\text{Hans, isst, Käsebrot, ein}\}$

$N = \{S, NP, VP, V, N, Det\}$; Startsymbol: S

Produktionsregeln:

$S \rightarrow NP \ VP$

$V \rightarrow \text{isst}$

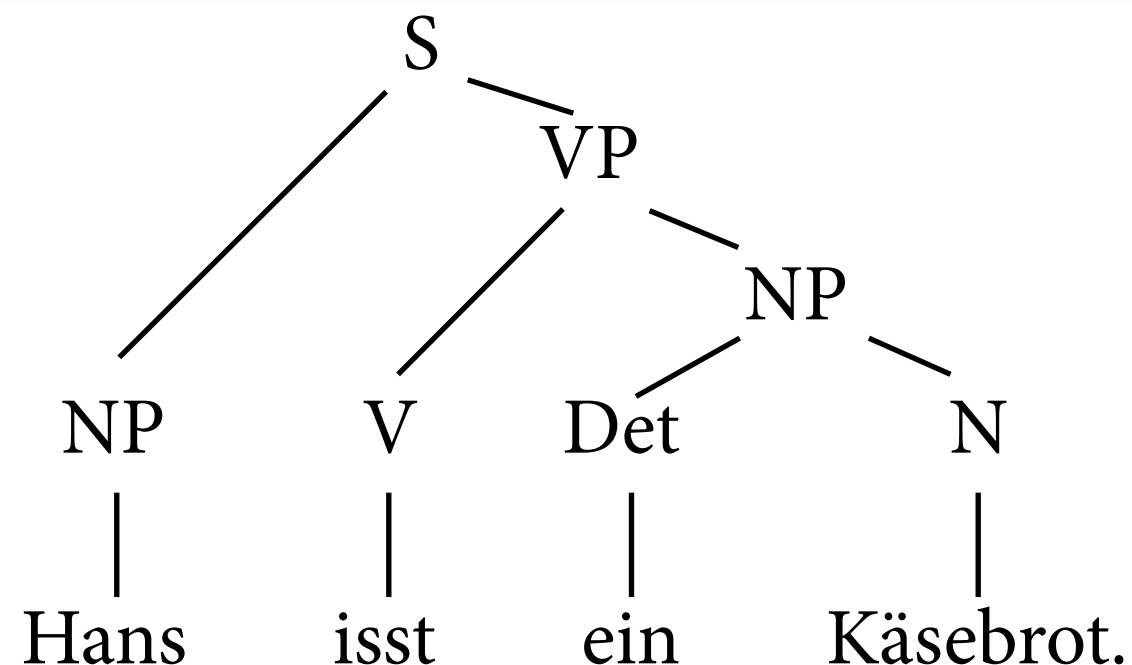
$Det \rightarrow \text{ein}$

$NP \rightarrow Det \ N$

$NP \rightarrow \text{Hans}$

$N \rightarrow \text{Käsebrot}$

$VP \rightarrow V \ NP$



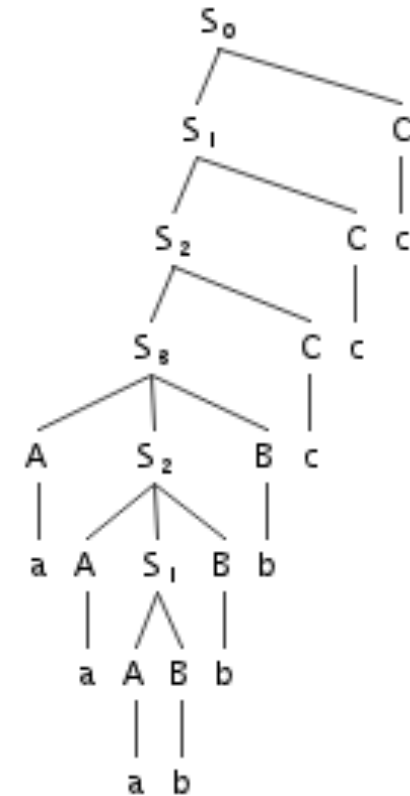
Beispiel: kfGs

- Kontextfreie Grammatiken (kfGs) sind ein *Grammatikformalismus*.
 - ▶ formales Werkzeug, mit dem man einzelne konkrete Grammatiken aufschreiben kann
 - ▶ jede Grammatik beschreibt eine Sprache
- Passende *Grammatiktheorie*: Phrasenstrukturgrammatik.
 - ▶ Ausdrücke bestehen aus kleineren Ausdrücken.
 - ▶ Hierarchie wird in PS-Baum dargestellt.

Zusammenhang

- GF muss erlauben, alle syntaktischen Strukturen zu beschreiben, die GT erfordert.
- Zu jeder relevanten Sprache soll man mit GF eine Grammatik schreiben können, die *für jeden Ausdruck* die richtigen Analysen beschreibt.
- *Expressivität* eines Grammatikformalismus:
Welche Klasse von Sprachen kann er beschreiben?

Expressivität



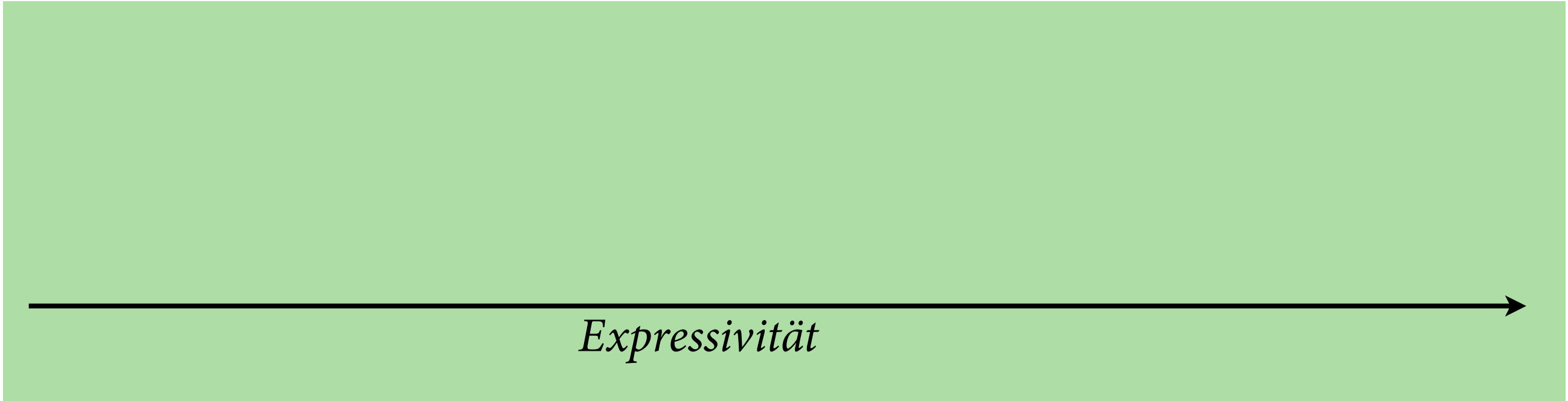
- Zu jedem einzelnen String gibt es einen PS-Baum, der mit irgendeiner kfG beschrieben werden kann.
- Aber die Sprache $a^n b^n c^n$ ist nicht kontextfrei.

Ansprüche an GF in der CL

- In der Computerlinguistik stellen wir an einen GF noch weitere Ansprüche.
 - ▶ Für große Sprachen (z.B. Deutsch) soll es nicht nur möglich, sondern auch *bequem* sein, Grammatik zu schreiben.
 - ▶ Verfügbarkeit von *effizienten* Parsingalgorithmen.
 - ▶ Verfügbarkeit von *statistischen Modellen*, die aus Korpusdaten trainiert werden können.

Natürliche Sprachen in der Chomsky-Hierarchie

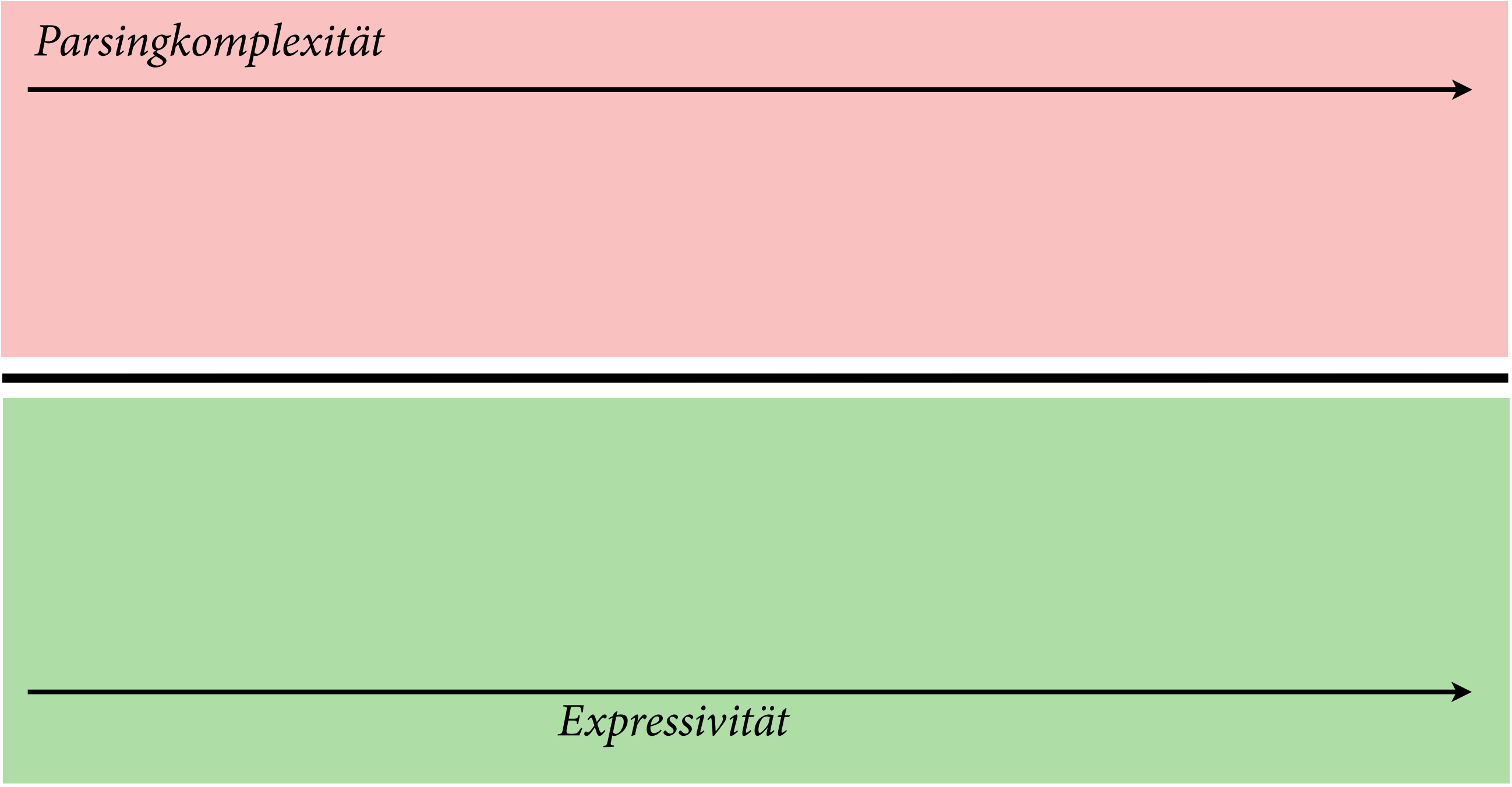
Natürliche Sprachen in der Chomsky-Hierarchie



Expressivität

Natürliche Sprachen in der Chomsky-Hierarchie

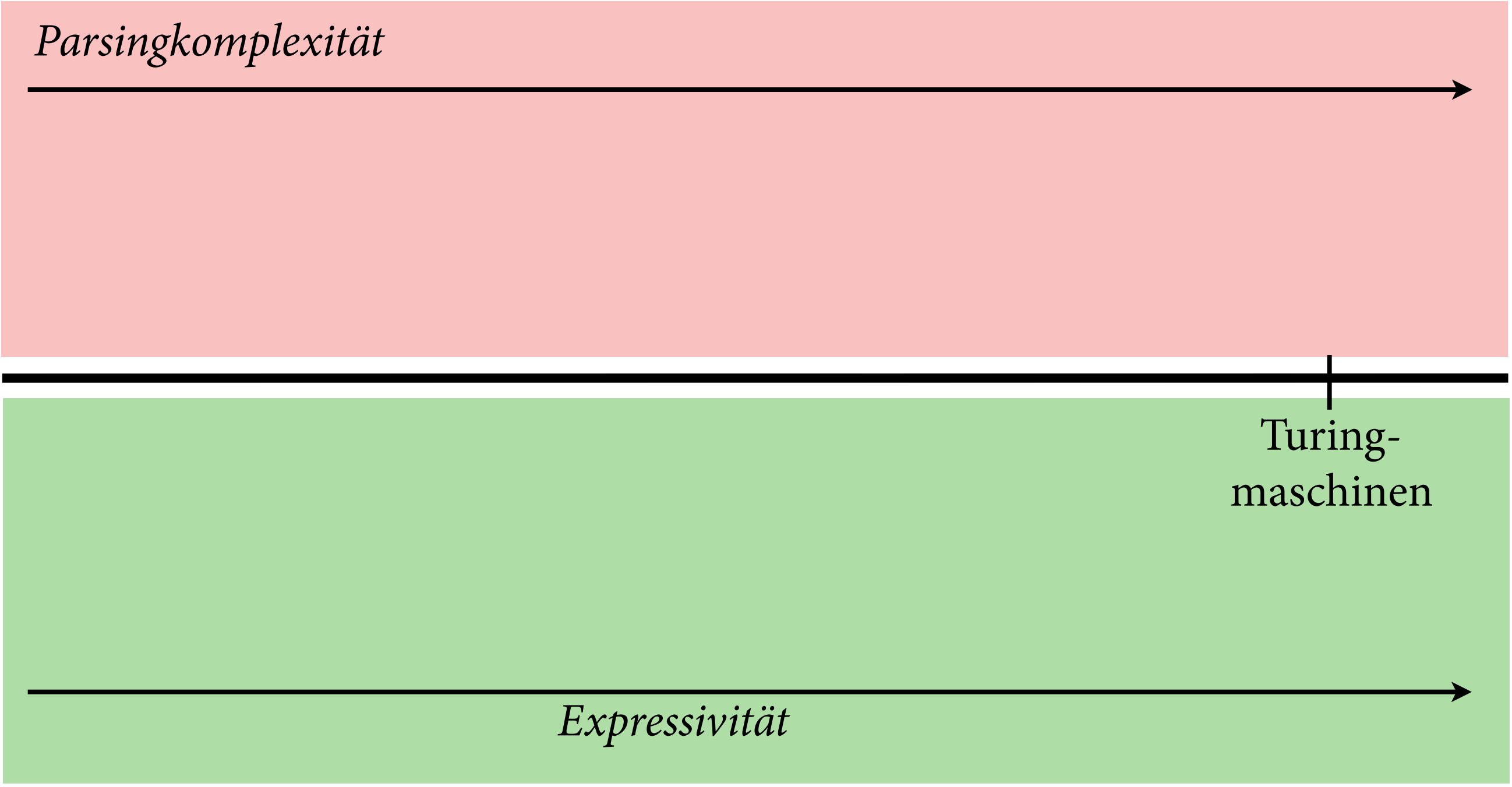
Parsingkomplexität



Expressivität

Natürliche Sprachen in der Chomsky-Hierarchie

Parsingkomplexität



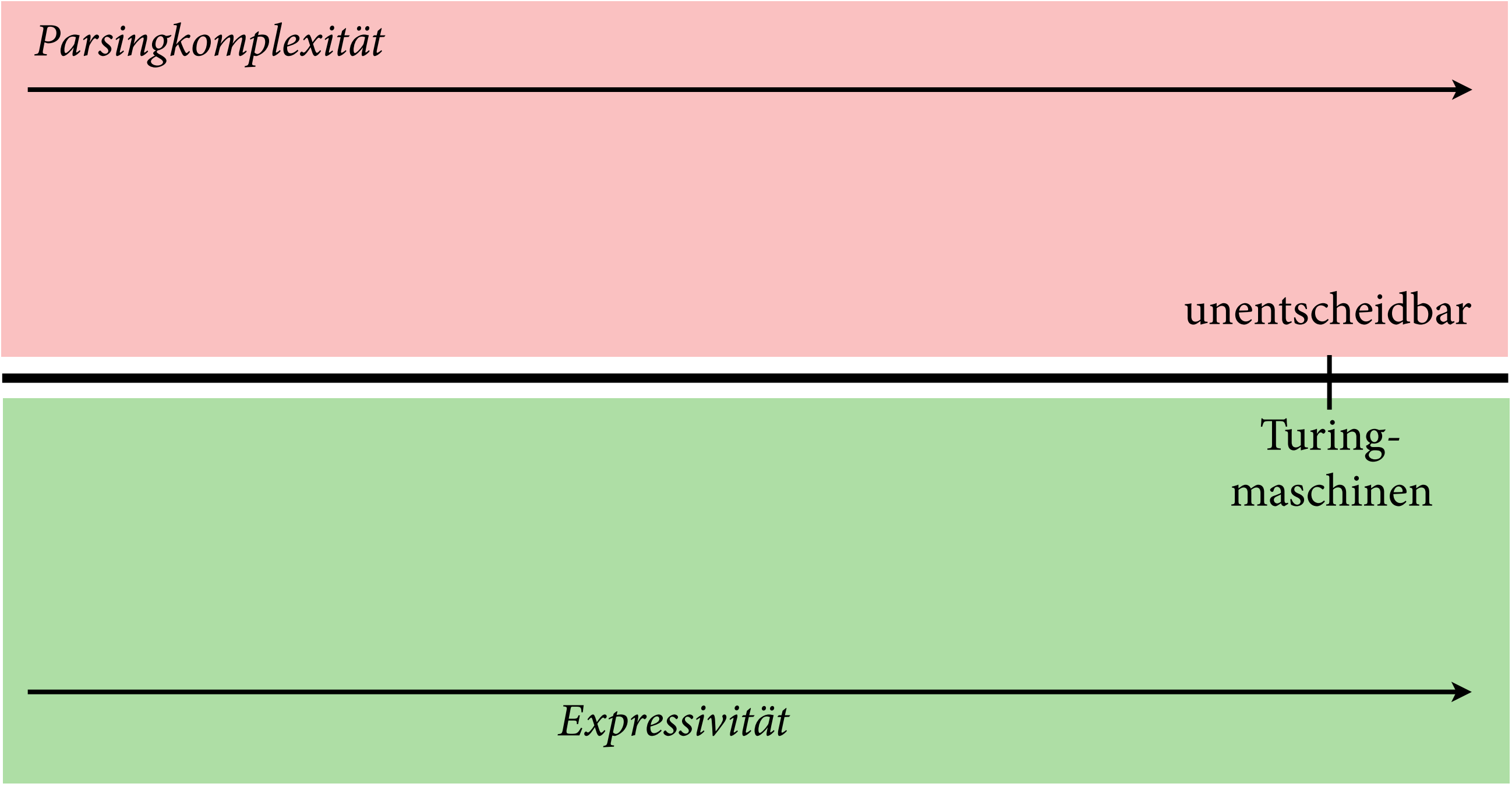
The diagram consists of two horizontal rectangular regions. The top region is light red and contains the text 'Parsingkomplexität' at its top left. Below this text is a horizontal black line with an arrow pointing to the right. The bottom region is light green and contains the text 'Expressivität' at its bottom center. Above this text is a horizontal black line with an arrow pointing to the right. A thick black horizontal line separates the two regions. On the right side of this separator line, there is a vertical tick mark pointing down into the green region, with the text 'Turing-maschinen' positioned to its right.

Turing-
maschinen

Expressivität

Natürliche Sprachen in der Chomsky-Hierarchie

Parsingkomplexität



The diagram consists of two horizontal rectangular regions. The top region is light red and contains the text 'Parsingkomplexität' at the top left and 'unentscheidbar' at the bottom right. A horizontal arrow points from left to right across this region. The bottom region is light green and contains the text 'Expressivität' at the bottom center. A horizontal arrow points from left to right across this region. A thick black horizontal line separates the two regions. A vertical black line segment crosses this horizontal line, with the text 'Turing-maschinen' positioned to its right.

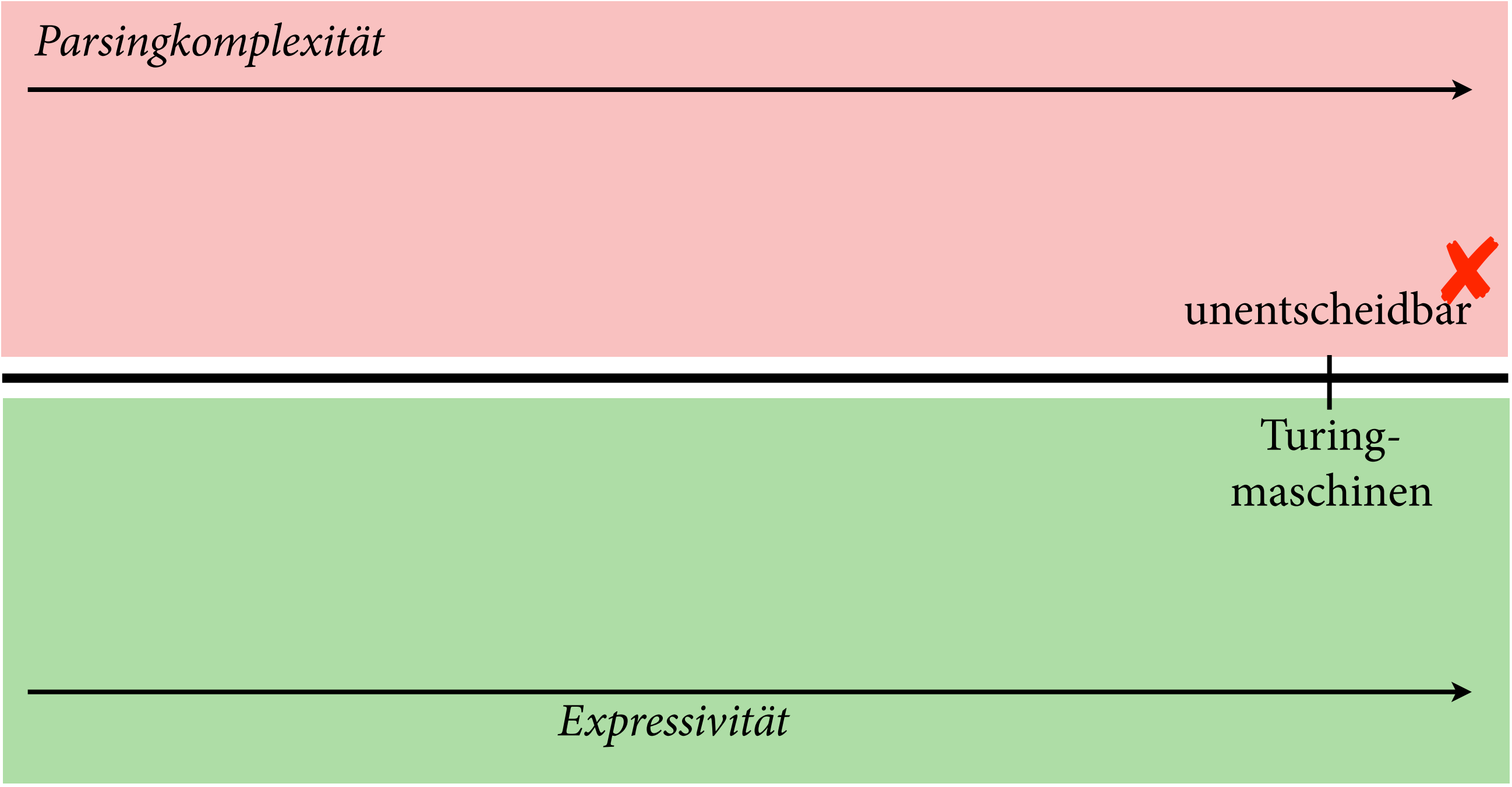
unentscheidbar

Turing-
maschinen

Expressivität

Natürliche Sprachen in der Chomsky-Hierarchie

Parsingkomplexität



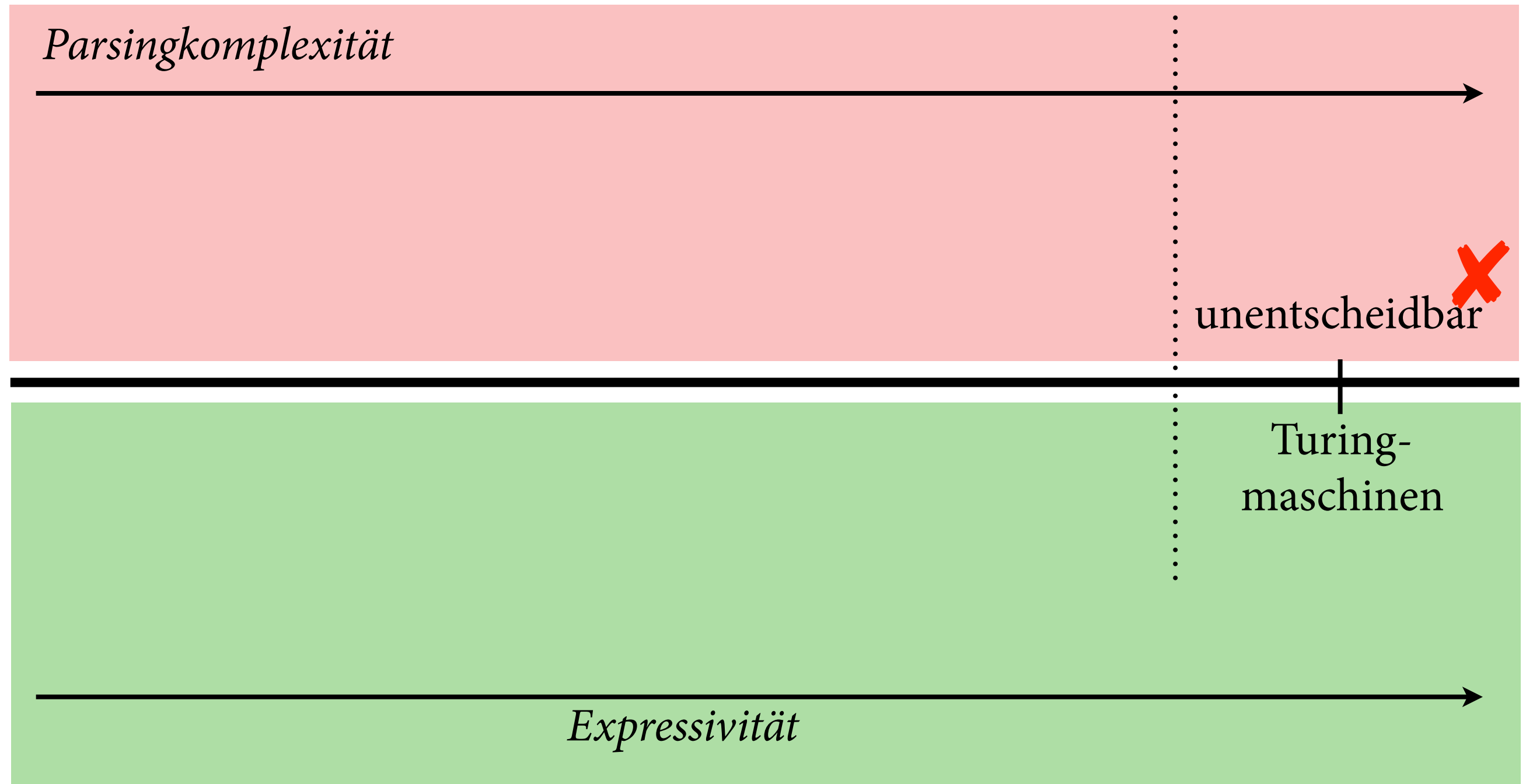
The diagram consists of two horizontal bars. The top bar is light red and contains the text 'Parsingkomplexität' at its left end and 'unentscheidbar' with a red 'X' at its right end. The bottom bar is light green and contains the text 'Expressivität' at its left end. A thick black horizontal line separates the two bars. A vertical black line with a tick mark at the bottom intersects the thick horizontal line, and this intersection point is labeled 'Turing-maschinen'.

unentscheidbar 

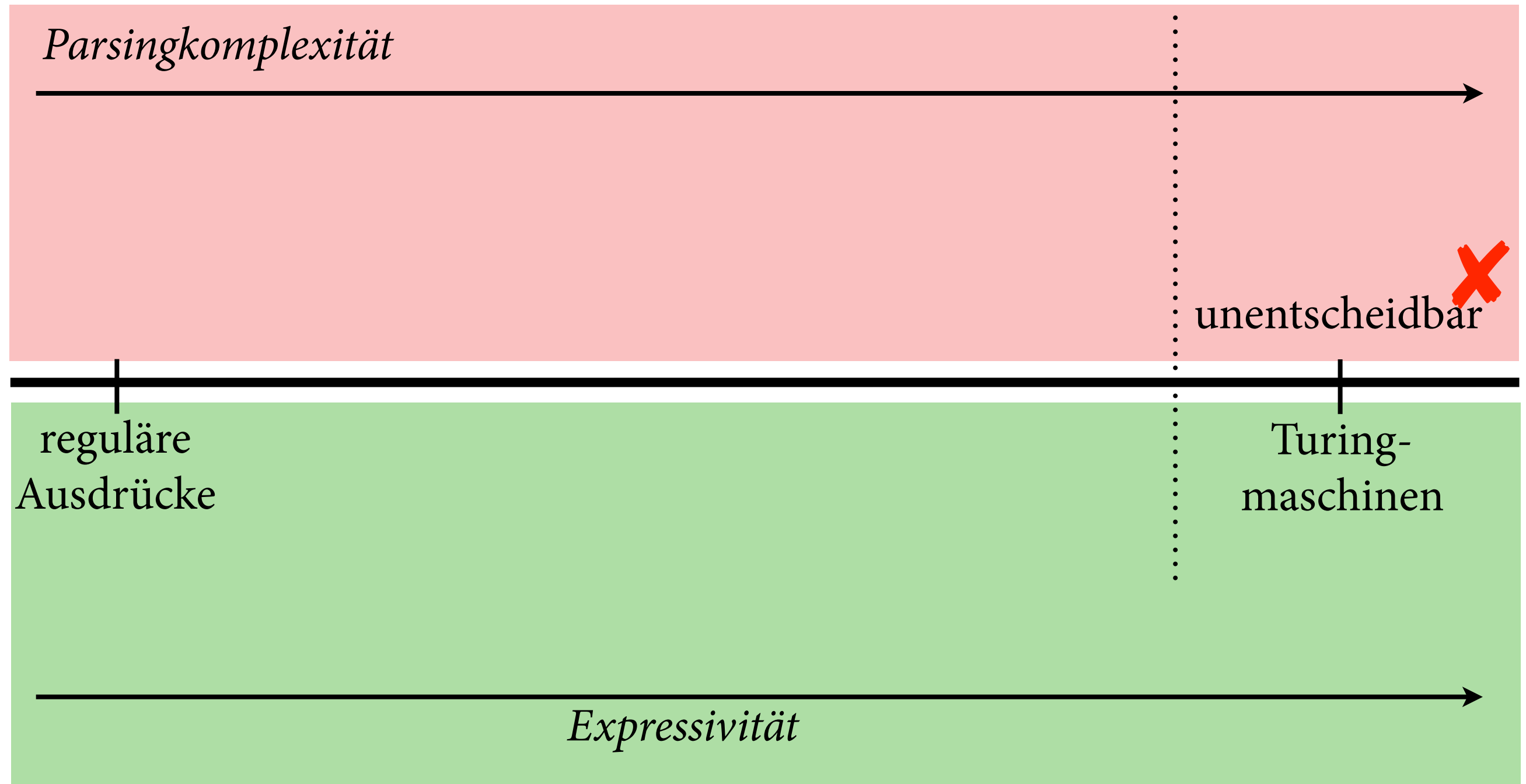
Turing-
maschinen

Expressivität

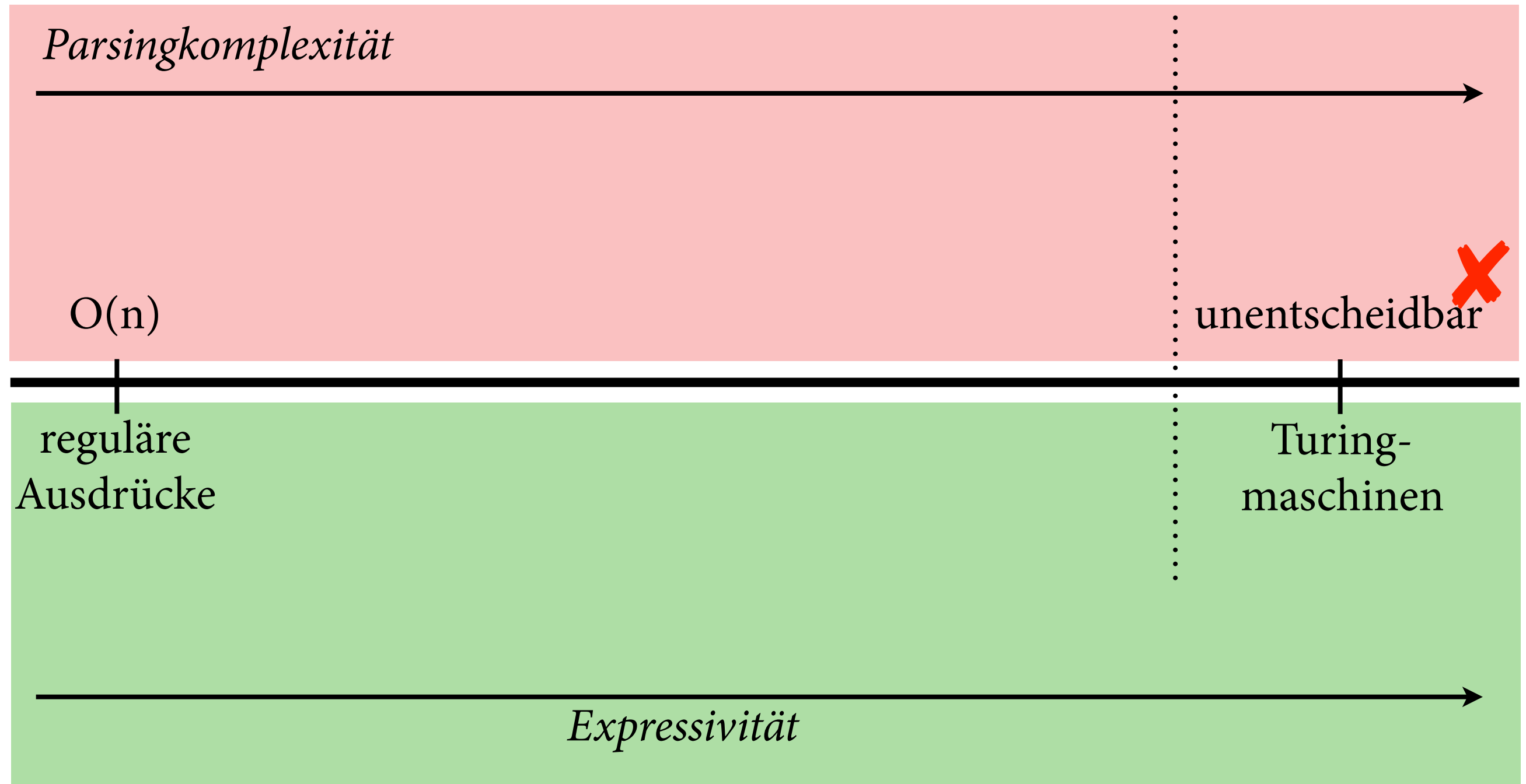
Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



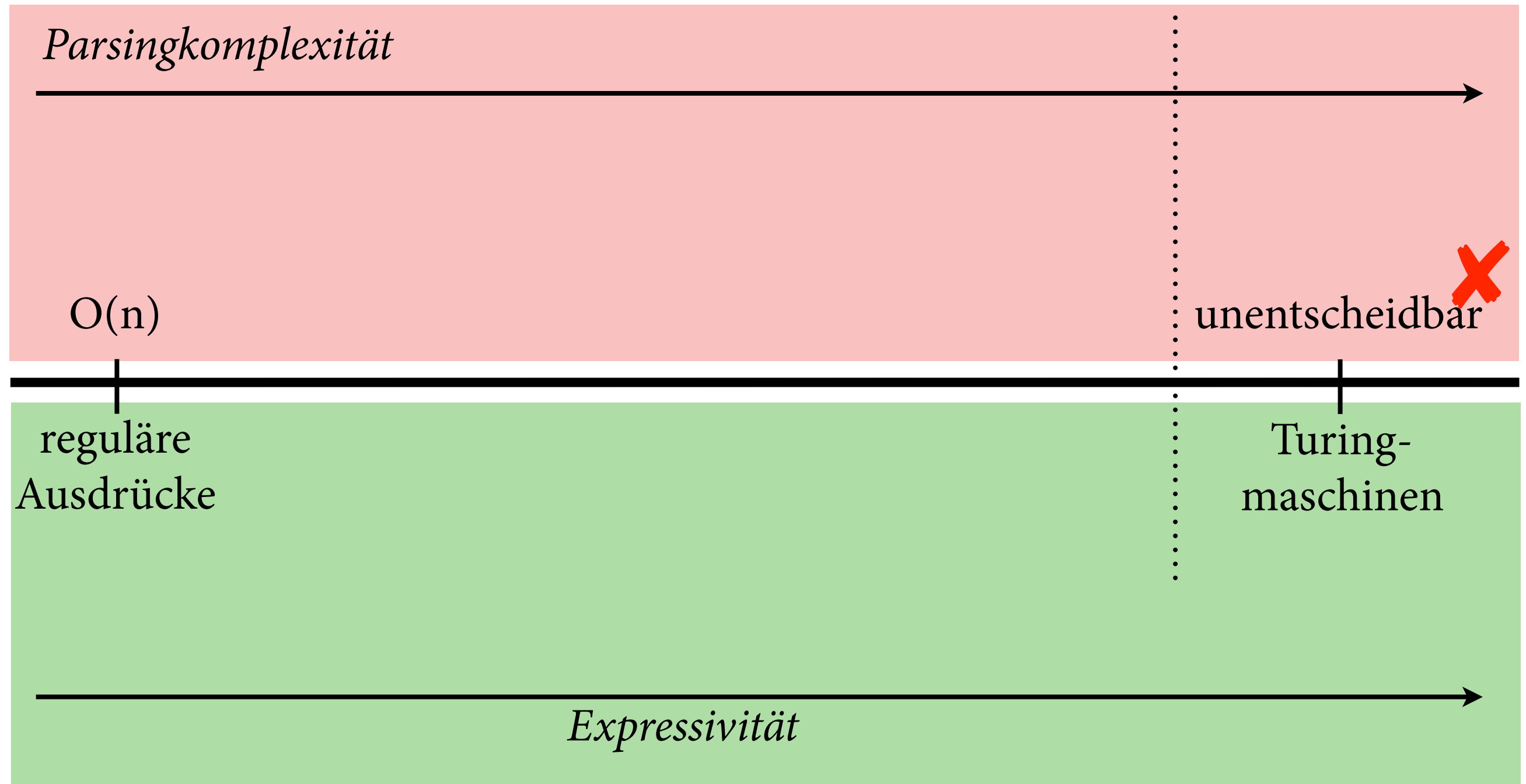
NL sind nicht regulär

... denn sie erlauben *Rekursion*.

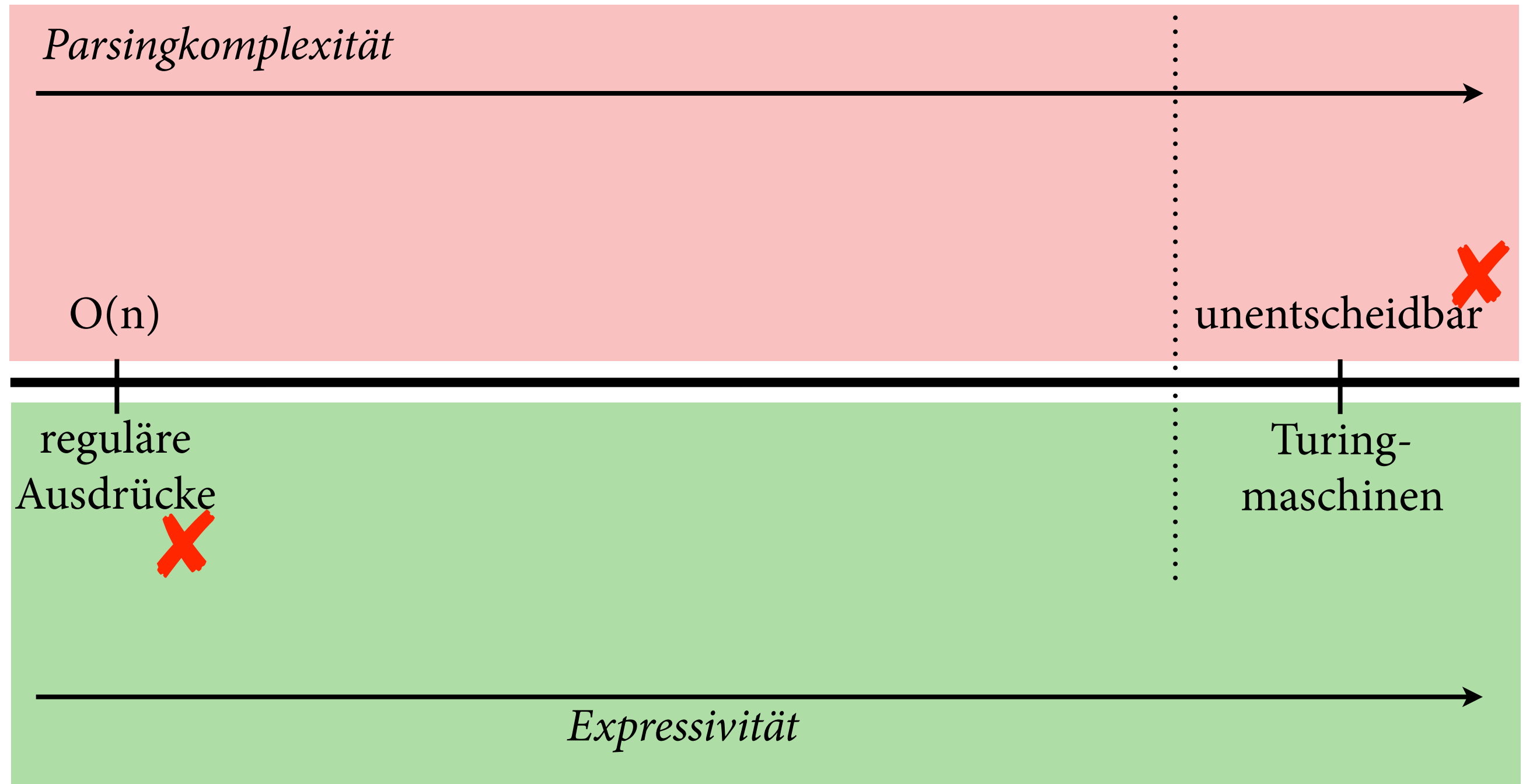
Der Hund					bellt.
a					b
Der Hund, der die Katze				jagt, bellt.	
a		a		b	b
Der Hund, der die Katze, die die Maus sucht, jagt, bellt.					
a		a		a	b
				b	b
					b

Geht für beliebig viele a's und b's. Aber $a^n b^n$ ist keine reguläre Sprache. Also ist Deutsch nicht regulär.

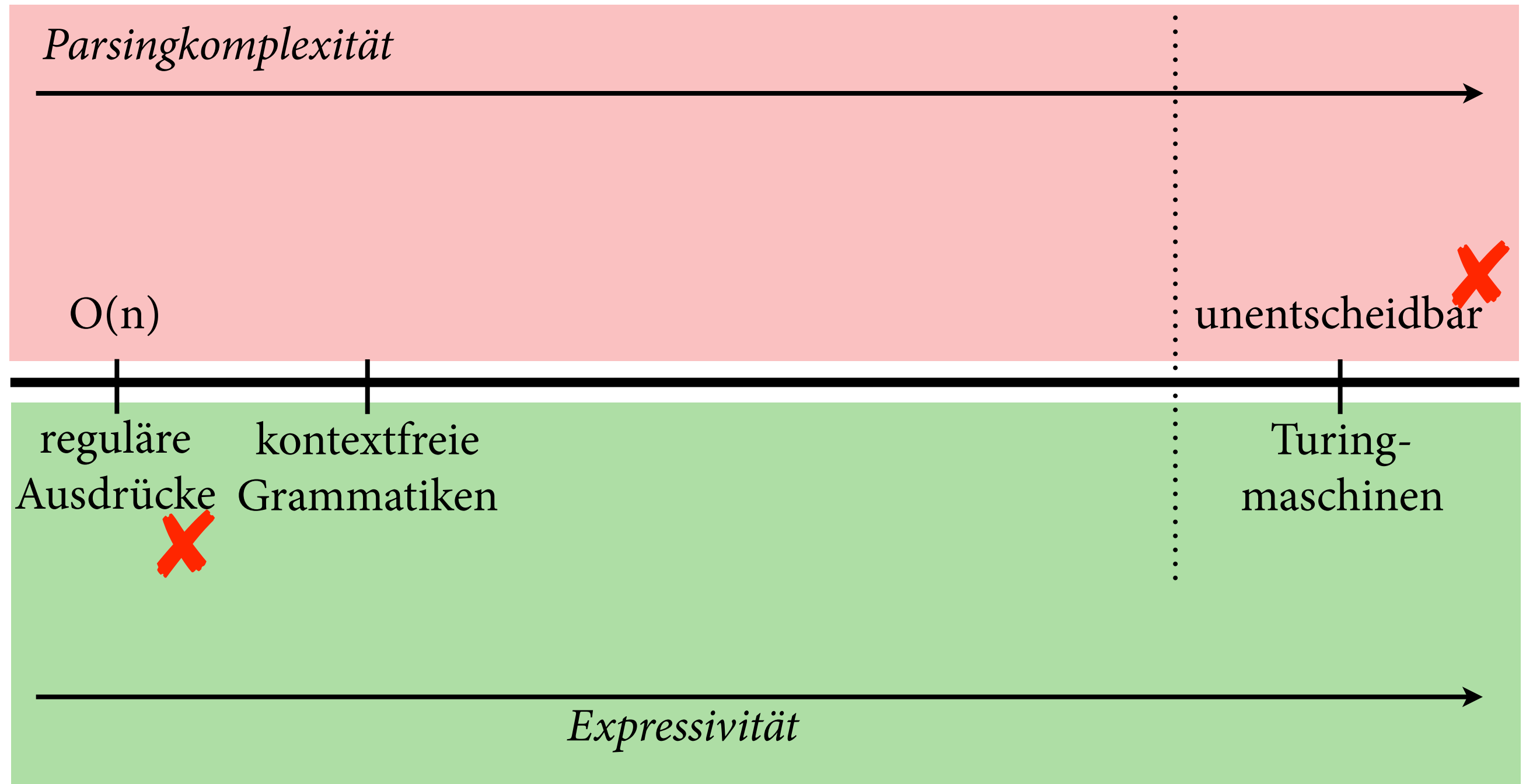
Natürliche Sprachen in der Chomsky-Hierarchie



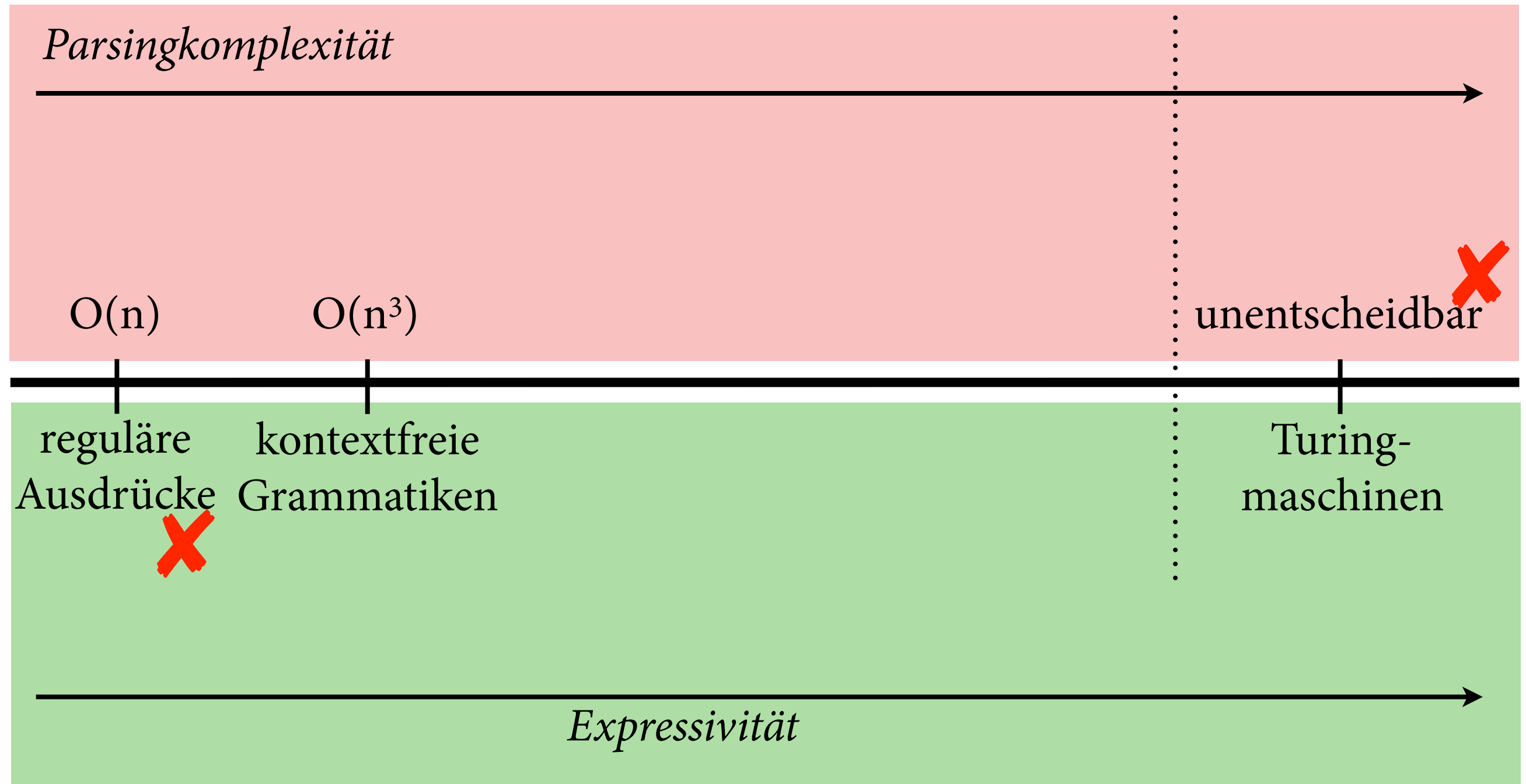
Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



NL sind nicht kontextfrei

... denn manche erlauben *cross-serial dependencies*.

Jan säit das mer es huus haend wele aastriiche.

Jan säit das mer d'chind es huus haend wele laa aastriiche.

Jan säit das mer em Hans es huus haend wele hälfe aastriiche.

Jan säit das mer d'chind em Hans es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans d'chind es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans em Sepp es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans es huus haend wele laa hälfe aastriiche.

(Shieber 1985 zu Zürcherdeutsch, sollte jeder Computerlinguist gelesen haben)

NL sind nicht kontextfrei

... denn manche erlauben *cross-serial dependencies*.

Jan säit das mer es huus haend wele aastriiche.

Jan säit das mer d'chind es huus haend wele laa aastriiche.

Jan säit das mer em Hans es huus haend wele hälfe aastriiche.

Jan säit das mer d'chind em Hans es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans d'chind es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans em Sepp es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans es huus haend wele laa hälfe aastriiche.

$\Rightarrow w \text{ a}^m \text{ b}^n x \text{ c}^m \text{ d}^n y$

(Shieber 1985 zu Zürcherdeutsch, sollte jeder Computerlinguist gelesen haben)

NL sind nicht kontextfrei

... denn manche erlauben *cross-serial dependencies*.

Jan säit das mer es huus haend wele aastriiche.

Jan säit das mer d'chind es huus haend wele laa aastriiche.

Jan säit das mer em Hans es huus haend wele hälfe aastriiche.

Jan säit das mer d'chind em Hans es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans d'chind es huus haend wele laa hälfe aastriiche.

* Jan säit das mer em Hans em Sepp es huus haend wele laa hälfe aastriiche.

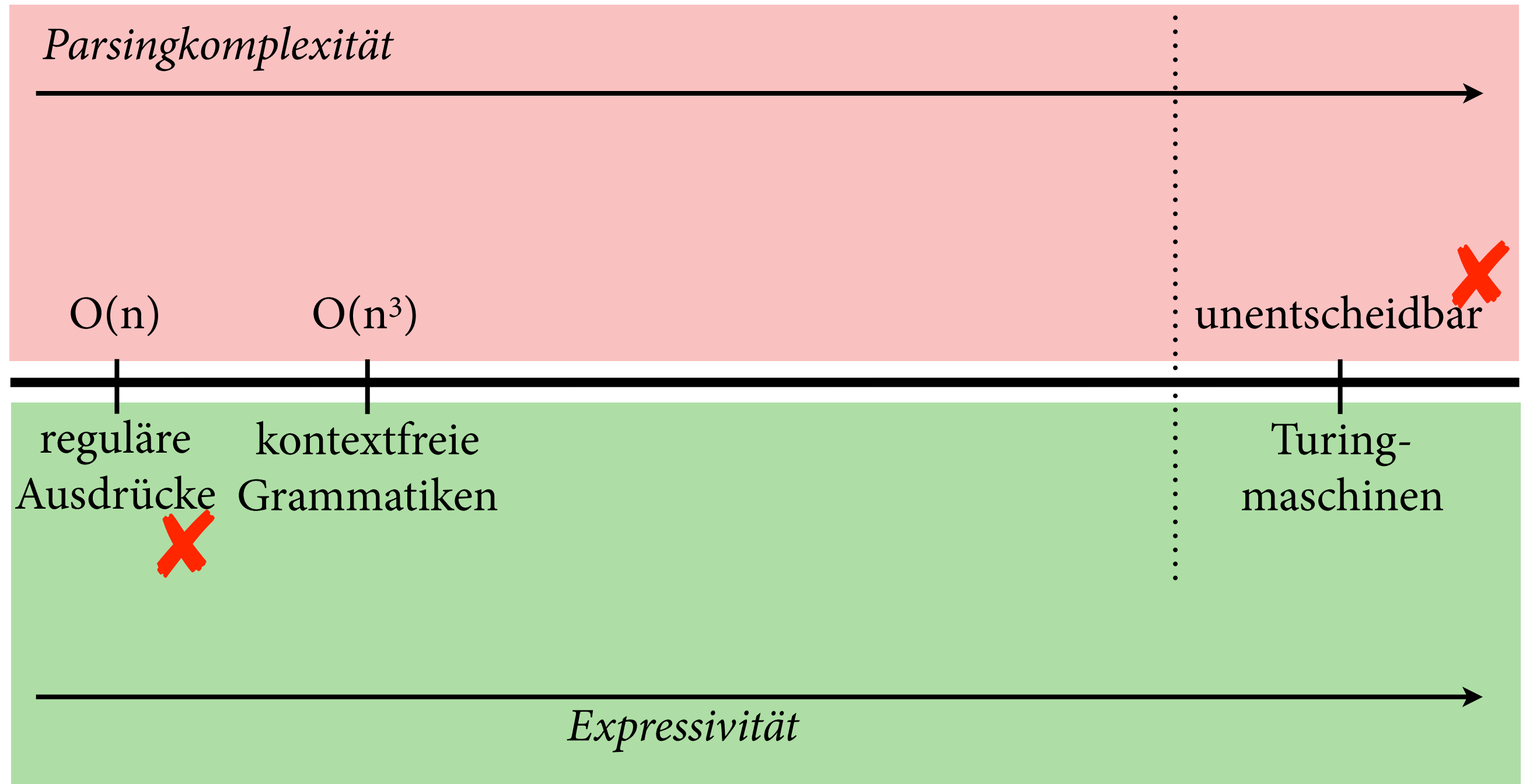
* Jan säit das mer em Hans es huus haend wele laa hälfe aastriiche.

$\Rightarrow w a^m b^n x c^m d^n y$

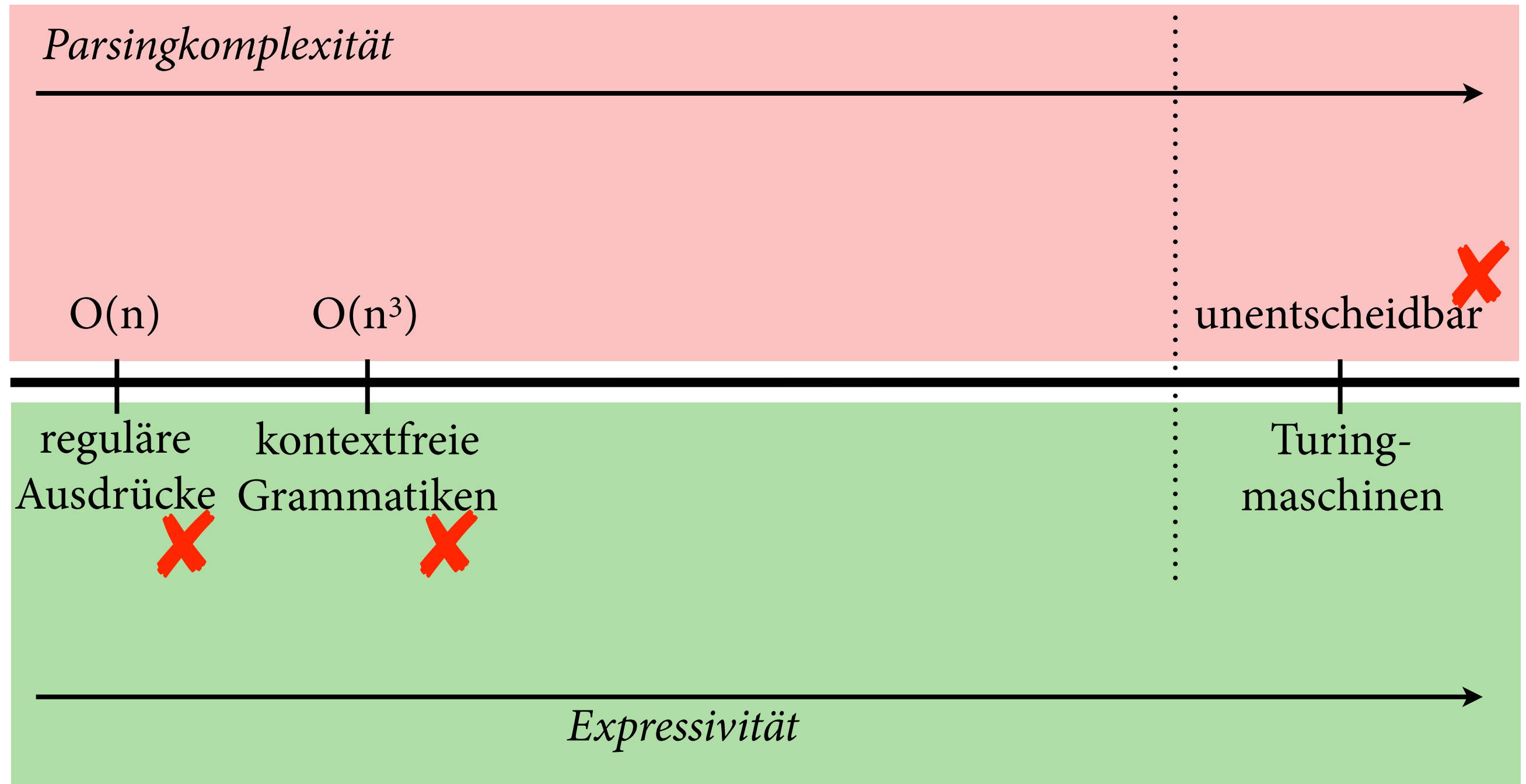
$\Rightarrow$ nicht kontextfrei

(Shieber 1985 zu Zürcherdeutsch, sollte jeder Computerlinguist gelesen haben)

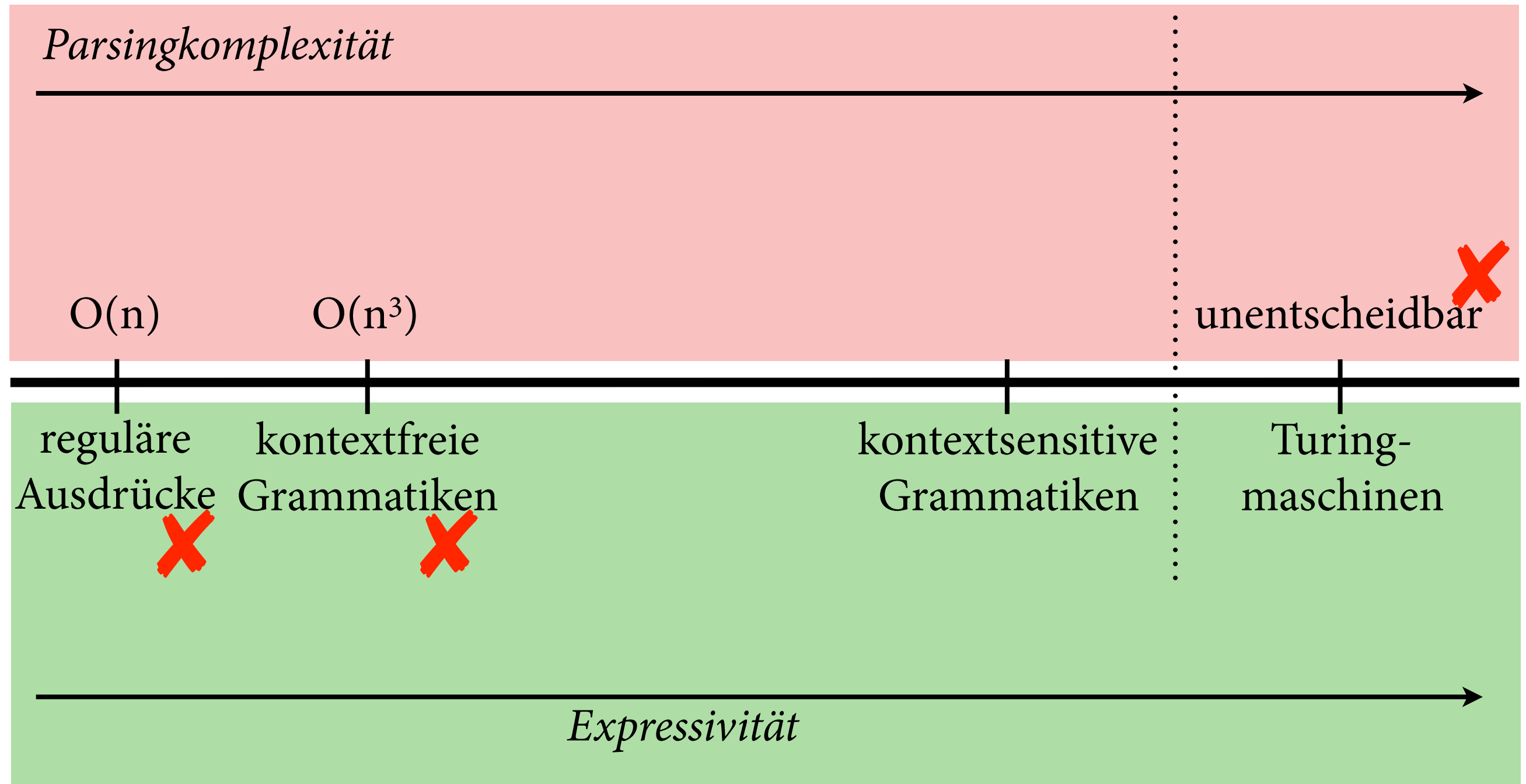
Natürliche Sprachen in der Chomsky-Hierarchie



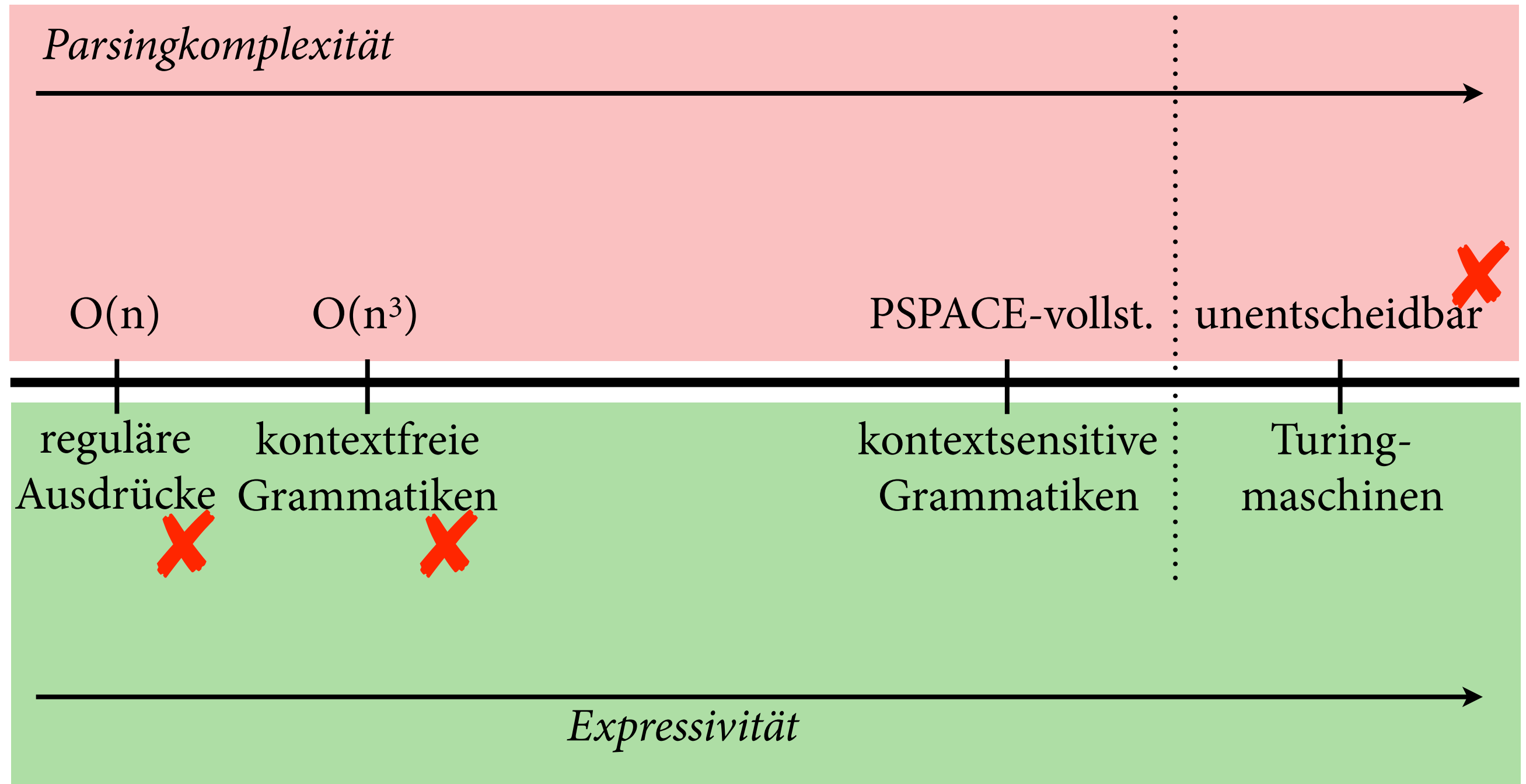
Natürliche Sprachen in der Chomsky-Hierarchie



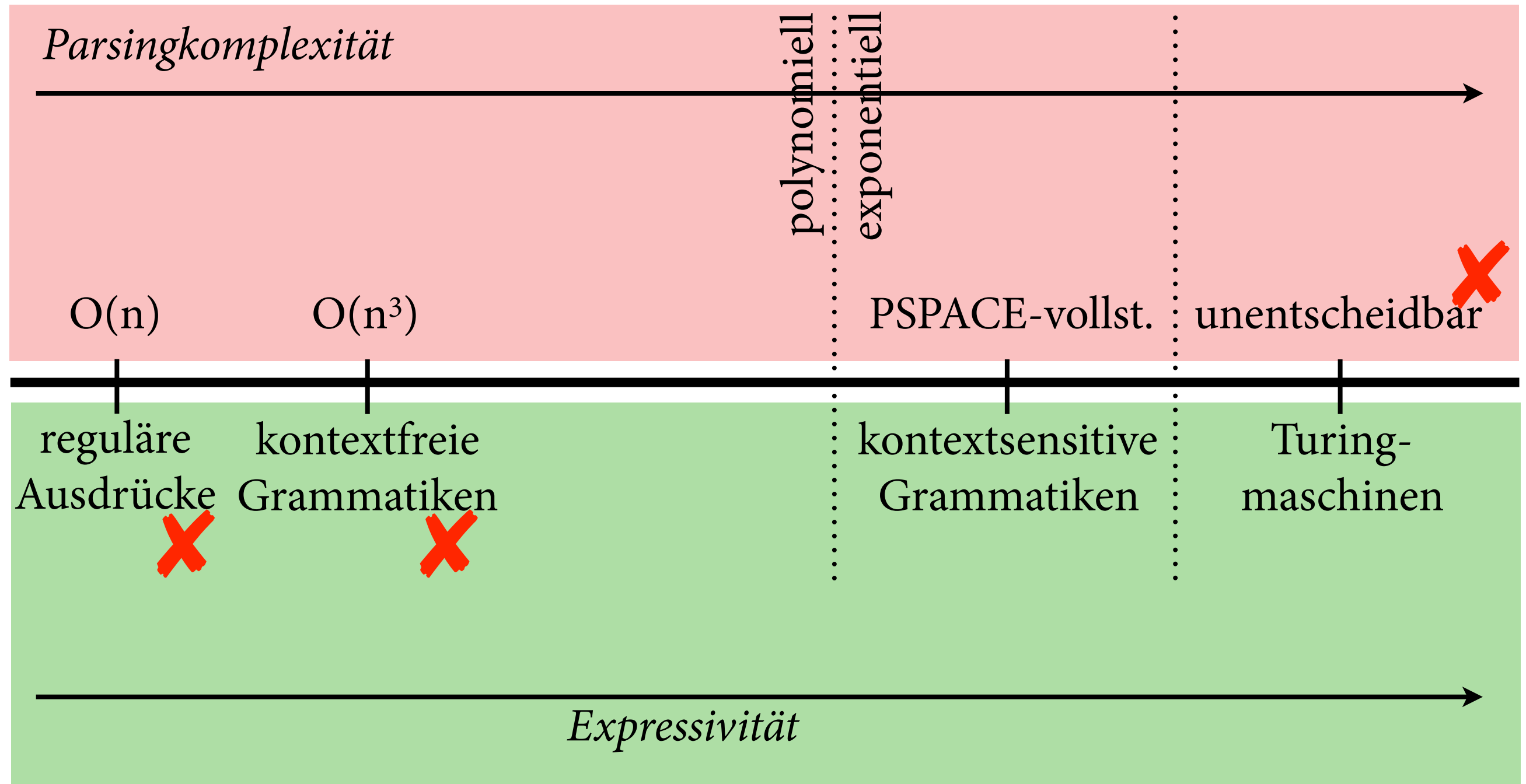
Natürliche Sprachen in der Chomsky-Hierarchie



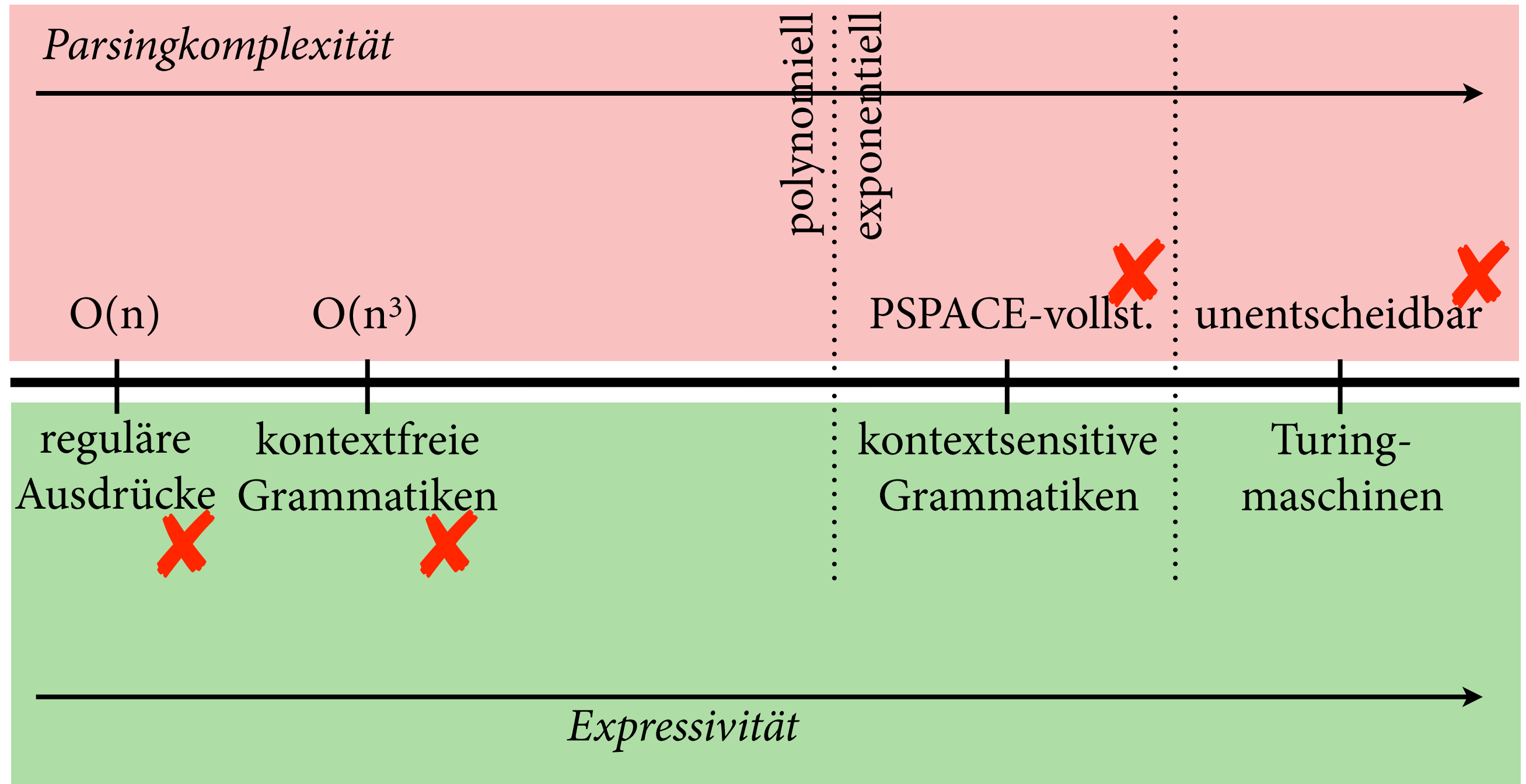
Natürliche Sprachen in der Chomsky-Hierarchie



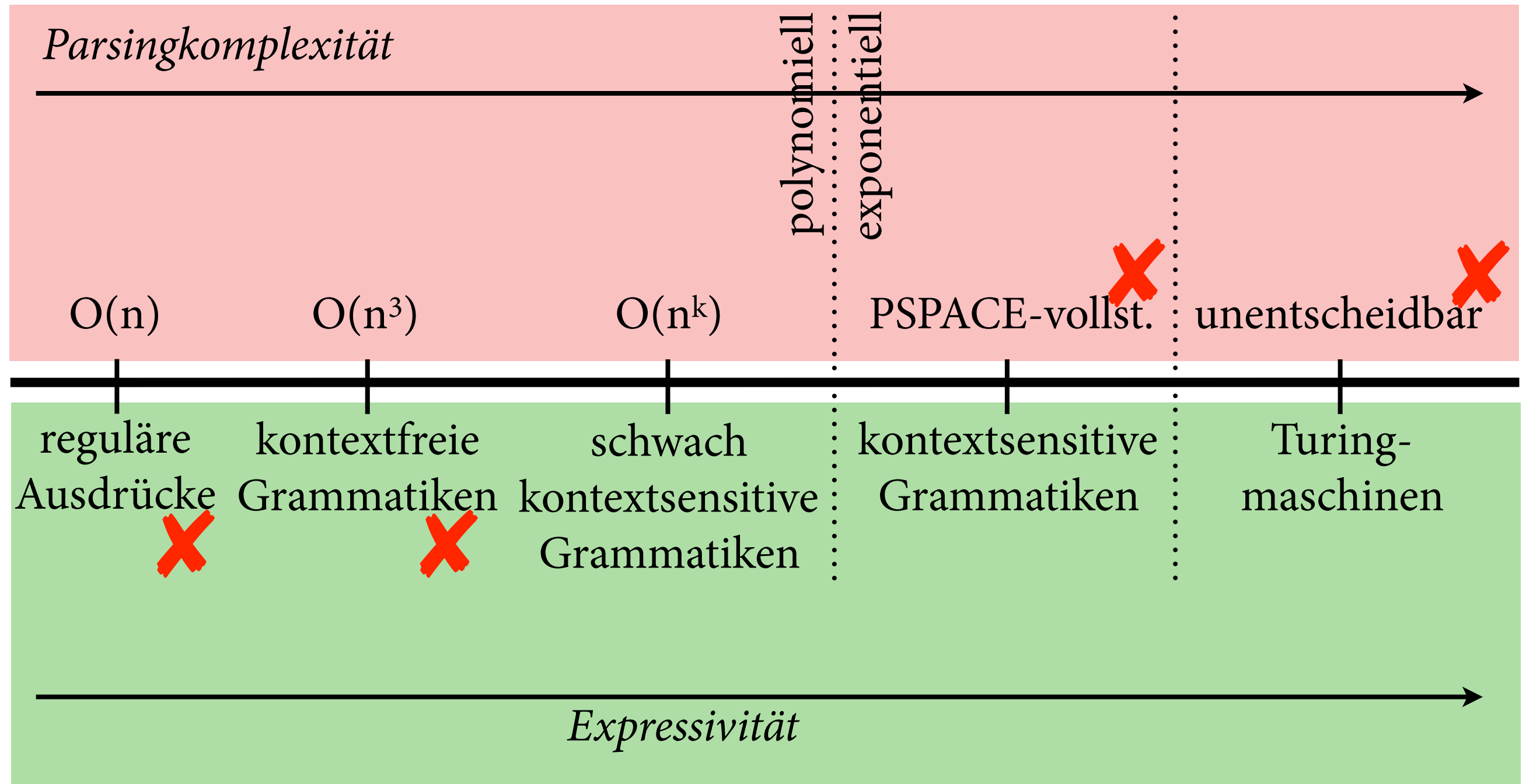
Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



Natürliche Sprachen in der Chomsky-Hierarchie



Kompetenz vs. Performanz

- Argumente basieren auf beliebig langen Sätzen, die kein normaler Mensch so sagt.
 - ▶ “die ein normaler Mensch sagt” = Performanz
- Trotzdem sind sie grammatisch korrekt.
 - ▶ = Kompetenz
- Performanzargument hat Charme, aber:
 - ▶ Wollen wir argumentieren, dass NL endlich sind?
 - ▶ Expressivere Formalismen auf jeden Fall angemessenere Werkzeuge zur Grammatikentwicklung.

Einige Probleme mit kfGs

- KfGs sind nicht nur wegen Expressivität kein idealer Formalismus für NL.
- Außerdem Probleme beim Erfassen linguistischer Verallgemeinerungen.
- Wir betrachten:
 - ▶ Kongruenz und Subkategorisierung
 - ▶ Scrambling

Subkategorisierung

- Verschiedene Verben stellen verschiedene Ansprüche daran, was für Komplemente sie brauchen (*subkategorisieren für diese*).

Hans schläft.	NP
Hans isst ein Käsebrot.	NP NP
Hans gibt dem Jungen ein Buch.	NP NP NP
Hans glaubt, dass es regnet.	NP S

- Subkategorisierte Komplemente heißen der *Subkategorisierungsrahmen* oder die *Valenz* des Verbs.

Kongruenz

- Verb gibt außerdem Kasus und evtl. Numerus der Komplemente vor.

Hans schläft.

NP_n

Hans isst ein Käsebrot.

NP_n NP_a

Hans gibt dem Jungen ein Buch.

NP_n NP_d NP_a

Hans glaubt, dass es regnet.

NP_n S

- Grammatik muss dafür sorgen, dass Valenz und Kongruenz eingehalten werden.

In einer kfG

Version 1: Eine Kategorie für Verben.

$VP \rightarrow V$

$VP \rightarrow V NP$

$VP \rightarrow V NP NP$

$V \rightarrow \text{schläft}$

$V \rightarrow \text{isst}$

$V \rightarrow \text{gibt}$

... erzwingt aber nicht Valenz.

In einer kfG

Version 2: Differenzierte Kategorien für jede Valenz.

VP $\rightarrow$ IV

VP $\rightarrow$ TV NP

VP $\rightarrow$ DV NP NP

IV $\rightarrow$ schläft

TV $\rightarrow$ isst

DV $\rightarrow$ gibt

... erfasst aber noch nicht Kongruenz.

In einer kfG

Version 3: Differenzierte Kategorien für jede Valenz und Kasus.

VP → IV

VP → TV NP_a

VP → DV NP_d NP_a

IV → schläft

TV → isst

DV → gibt

NP_a → den N_{masc}

NP_d → dem N_{masc}

NP_a → die N_{fem}

NP_d → der N_{fem}

NP_a → das N_{neut}

NP_d → dem N_{neut}

Das Problem

- Allein für Kasus brauchen wir drei Nichtterminale und sechs NP-Regeln.
 - ▶ x2 für Numerus
 - ▶ für Adjektiv-Nomen-Kongruenz x3 für Genus
- Problem aus linguistischer Sicht:
Übergreifende Zusammenhänge nicht dargestellt.
- Problem aus CL-Sicht:
viel zu große und nicht wartbare Grammatik.

Wald vor lauter Bäumen

- Valenz-Regeln sind alles Spezialfälle einer allgemeineren Regel:

$$V[\text{subcat}_1, \dots, \text{subcat}_n] \rightarrow \\ V[\text{subcat}_2, \dots, \text{subcat}_n] \quad \text{subcat}_1$$

- Ebenso für Kongruenz:
“Kasus von subcat_i muss zu V passen.”
- Diese allgemeinen Regeln sind aber nicht mit kfG darstellbar.

Freie Wortstellung

- Sprachen unterscheiden sich in den Ausmaß, in dem Wortstellung variiert werden kann.
 - ▶ Englisch: sehr festgelegt
 - ▶ Latein: völlig frei
 - ▶ Deutsch: einigermaßen frei
- KfG-Regel legt aber Reihenfolge der Komplemente vollständig fest.

$VP \rightarrow DV \text{ NP}_d \text{ NP}_a$

Scrambling im Deutschen

- Lokale Umstellung der Verbkomplemente:

(weil) der Mann der Frau das Buch gibt

(weil) der Mann das Buch der Frau gibt

(weil) der Frau der Mann das Buch gibt

(weil) das Buch der Mann der Frau gibt

(weil) der Frau das Buch der Mann gibt

(weil) das Buch der Frau der Mann gibt

- Komplemente verschiedener Verben können gemischt werden:

(dass) der Detektiv dem Klienten den Verdächtigen
des Verbrechens zu überführen versprochen hat

(dass) des Verbrechens der Detektiv den Verdächtigen
dem Klienten zu überführen versprochen hat

Scrambling mit kfGs

- Lokales Scrambling mit kfG darstellbar, aber Grammatik wird exponentiell groß:

$VP \rightarrow NP_n NP_a NP_d V$

$VP \rightarrow NP_n NP_d NP_a V$

$VP \rightarrow NP_a NP_n NP_d V$

$VP \rightarrow NP_d NP_n NP_a V$

$VP \rightarrow NP_a NP_d NP_n V$

$VP \rightarrow NP_d NP_a NP_n V$

Variante: ID/LP-Grammatiken

- Lösungsversuch:
Trenne in kfG Beschreibung von
 - ▶ *Valenz* = immediate dominance = ID
 - ▶ *Wortstellung* = linear precedence = LP
- Regeln in ID/LP-Grammatik:
 - ▶ haben Form $A \rightarrow B_1, \dots, B_n$ und geben zunächst nur an, dass *A Kinder* $B_1, \dots, B_n$ im Baum haben muss
 - ▶ lineare Abfolge von $B_1, \dots, B_n$ kann durch *LP-Constraints* $B_i < B_j$ erzwungen werden

ID/LP-Grammatik für Scrambling

- Lokales Scrambling geht so:

$$\begin{aligned} &VP \rightarrow NP_n, NP_a, NP_d, V \\ &NP_n < V, NP_a < V, NP_d < V \end{aligned}$$

- Sechs kfG-Regeln in 1 ID/LP-Regel dargestellt.
- Problem: Wortproblem von ID/LP-Grammatiken ist NP-vollständig. Worst case: exponentielle Laufzeit.

Ablauf der Vorlesung

kontextfreie
Grammatiken

```
graph TD; A[kontextfreie Grammatiken] --> B[reichere Ableitungen]; A --> C[Features];
```

reichere Ableitungen

- Baumadjunktionsgrammatik (TAG)
- Kombinatorische Kategorialgrammatik (CCG)
- weitere schwach kontext-sensitive Formalismen

Features

- Lexikalisch-funktionale Grammatik (LFG)
- Head-driven PSG (HPSG)

Wiederkehrende Themen

- Wie stellt GF den Zusammenhang von Valenz und Wortstellung dar?
- Parsingalgorithmen und Effizienz
- Expressivität, auch im Vergleich mit anderen GFen
- Passende Wahrscheinlichkeitsmodelle

Literatur

- Auf jeden Fall herunterladen:
Müller (2010), *Grammatiktheorie*.
- Weitere nützliche Bücher:
 - ▶ TAG: Kallmeyer (2010), *Parsing beyond context-free grammars*.
 - ▶ CCG: Steedman (2000), *The syntactic process*.

Übungen

- Zu jedem GF wird es eine Übung geben (insgesamt 5 Übungen).
- Schwerpunkt der Übungen:
Selbst Grammatiken entwickeln.
- Tutorium: Theresa Schmidt
- Abschlussprojekt in letzter Sitzung. Zulassung:
 - ▶ mindestens 4 Übungen bearbeiten
 - ▶ mindestens 50% der möglichen Punkte aus diesen 4

Ressourcen im Internet

- Alle Folien für die Vorlesung mache ich auf der Kurs-Homepage verfügbar.
- Piazza-Forum (Link auf Kurs-Webseite):
 - ▶ Ankündigungen
 - ▶ Daten für Übungen
 - ▶ Nutzen Sie Forum für Rückfragen und Diskussionen!
- Übungen bitte auf Piazza abgeben.

Zusammenfassung

- Grammatikformalismus - Grammatik - Grammatiktheorie - Sprache
- NL im Spannungsfeld von Expressivität und Parsingeffizienz
- Einige Probleme mit kfG; Lösungsversuch mit ID/LP-Grammatiken.