

# E DIN 19461:2026-06 (D)

Erscheinungsdatum: 2026-05-01

## Sprachressourcen und Sprachtechnologie - Abgeleitete Textformate (ATF)

---

<b>Inhalt</b>	<b>Seite</b>
Vorwort .....	4
Einleitung .....	5
1 Anwendungsbereich.....	6
2 Normative Verweisungen .....	6
3 Begriffe .....	6
3.1 Grundsätzliches .....	6
3.2 Wissenschaftliche Methodiken .....	13
3.3 Rechtliche Rahmenbedingungen .....	17
3.4 Technische Verfahren .....	18
4 Anforderungen .....	21
4.1 Allgemeines .....	21
4.2 Allgemeine Anforderungen .....	21
4.3 Anforderungen an die Anreicherung von Daten .....	21
4.4 Anforderung an die Informationsreduktion .....	22
4.4.1 Allgemeines .....	22
4.4.2 Beibehalten .....	23
4.4.3 Löschen .....	23
4.4.4 Ersetzen .....	23
4.4.5 Randomisieren .....	23
4.5 Anforderung an die Kombination mehrerer ATFs .....	24
4.6 Anforderung an die Veröffentlichung von Daten in einem ATF .....	24
5 Erzeugung von abgeleiteten Textformaten.....	24
6 Anforderung an die Dokumentation des ATFs bei der Archivierung und Weitergabe.....	26
6.1 Allgemeines .....	26
6.2 Allgemeine Anforderungen an die Dokumentation des ATFs.....	26
6.3 Inhalt der Dokumentation.....	26
6.4 Buchführung über archivierte und weitergegebene ATFs.....	27
6.5 Verknüpfung der Dokumentation zu ATFs mit Metadaten.....	27
Anhang A (informativ) Elemente/Einheiten .....	28
A.1 Struktur .....	28
A.2 Text (ursprünglich oder verändert) .....	28
A.3 Annotation (exemplarisch / dynamisch erweiterbar).....	29
Anhang B (informativ) Beispiele abgeleiteter Textformate .....	30
B.1 Tokenbasierte ATFs.....	30
B.1.1 Term-Dokument-Matrix bzw. Bag of Words .....	30
B.1.2 Maskierung von Tokens .....	30
B.1.3 Segmentweise Aufhebung der Sequenzinformation.....	30
B.1.4 Selektiv reduzierte Information über einzelne Tokens.....	31
B.1.5 n-Gramm .....	31
B.2 Vektorbasierte ATFs.....	32
B.2.1 Wort-Embeddings .....	32
B.2.2 Kontextualisierte Embeddings .....	32
B.3 Beispiele für Verwendung von ATFs .....	33
B.3.1 CORDE: Term-Dokument-Matrix/Bag of Words.....	33
B.3.2 Mark Davies' Corpora: Maskierung von Tokens .....	33
B.3.3 CoNNSA im TextGrid Repository: Segmentweise Aufhebung der Sequenzinformation.....	33
B.3.4 American Drama Corpus: Selektiv reduzierte Information über einzelne Tokens .....	33
B.3.5 CorpusMasker / TüBa-D/Z.....	34
B.3.6 Google ngrams: n-Gramme.....	34
B.3.7 Hathi Trust Extracted Features: Kombination mehrerer Verfahren.....	34
Literaturhinweise .....	35