
Kajian 10 Jenis Kata (Wortarten) Pada Teks Bahasa Jerman “Das Gewitter” Dengan Pendekatan Komputasi Linguistik

Dea Enjelina Sipayung¹, Mika Risty Silaban², Rachel Inda Yemima Purba³, Hesti Maya Sari Saragih⁴

^{1,2,3,4}Universitas Negeri Medan

denjelina635@gmail.com¹, mikasilaban2233132039@mhs.unimed.ac.id²,

rachelpurba2003@gmail.com³, hestimaya.2233132039@mhs.unimed.ac.id⁴

ABSTRACT; *This research focuses on identifying ten word classes (Wortarten) and their roles in the German historical narrative text "Das Gewitter" by Ernst H. Gombrich, taken from the book Eine kurze Weltgeschichte für junge Leser. Employing a computational linguistics approach, this study applies the Part-of-Speech (POS) Tagging method based on the Stuttgart-Tübingen Tagset (STTS). The data analyzed consists of approximately 250 tokens that were manually annotated into the following categories: Nouns, Verbs, Articles, Pronouns, Prepositions, Adjectives, Conjunctions, Adverbs, Particles, and Numerals. The results show that Nouns dominate with a percentage of 27.2%, followed by Verbs (20.8%) and Articles (15.2%). This pattern reflects the high referential density characteristic of historical narrative texts. Furthermore, several computationally relevant linguistic phenomena were also identified, including word-class ambiguity, separable verbs (trennbare Verben), contracted preposition-article forms (Verschmelzungsform), as well as challenges in Named Entity Recognition (NER) for proper names.*

Keywords: *Wortarten, POS Tagging, STTS, Computational Linguistics, German Language.*

ABSTRAK; Penelitian ini berfokus pada identifikasi sepuluh kelas kata (Wortarten) serta perannya dalam teks naratif sejarah berbahasa Jerman berjudul “Das Gewitter” karya Ernst H. Gombrich yang terdapat dalam buku Eine kurze Weltgeschichte für junge Leser. Dengan memanfaatkan pendekatan linguistik komputasional, penelitian ini menerapkan metode POS (Part-of-Speech) Tagging berdasarkan sistem tagset STTS (Stuttgart-Tübingen-Tagset). Data yang dianalisis terdiri dari sekitar 250 token yang telah melalui proses anotasi manual ke dalam kategori: Nomina, Verba, Artikula, Pronomina, Preposisi, Adjektiva, Konjungsi, Adverbia, Partikel, dan Numeralia. Hasil menunjukkan bahwa Nomina mendominasi dengan persentase 27,2%, diikuti oleh Verba (20,8%) dan Artikula (15,2%). Pola ini mencerminkan karakteristik teks sejarah yang padat akan rujukan (referential density). Lebih lanjut, ditemukan pula sejumlah fenomena kebahasaan yang relevan secara komputasional, seperti ambiguitas kelas kata, verba terpisah (trennbare Verben), bentuk kontraksi preposisi-artikula (Verschmelzungsform), serta tantangan dalam pengenalan entitas bernama (Named Entity Recognition) pada nama diri.

Kata Kunci: Wortarten, POS Tagging, STTS, Linguistik Komputasional, Bahasa Jerman.

PENDAHULUAN

Linguistik komputasional adalah bidang kajian interdisipliner yang menggunakan pendekatan berbasis komputer untuk menelaah sistem struktur dan kandungan semantik bahasa manusia. Dalam ranah pemrosesan bahasa alami (Natural Language Processing/NLP), salah satu operasi yang paling mendasar adalah Part-of-Speech (POS) Tagging—sebuah prosedur anotasi yang menetapkan kategori gramatikal pada tiap satuan kata dalam sebuah ujaran atau teks (Jurafsky & Martin, 2023). Akurasi dalam POS tagging menjadi prasyarat utama bagi keberhasilan berbagai tugas NLP yang lebih kompleks, mulai dari penguraian sintaksis, pengenalan entitas bernama, hingga analisis opini dan sentimen.

Bahasa Jerman menjadi kasus yang secara linguistik menantang bagi sistem POS tagging karena memiliki sistem morfologi yang sangat elaboratif. Hal ini mencakup perubahan bentuk kata berdasarkan kasus gramatikal (Nominativ, Akkusativ, Dativ, dan Genitiv), pembagian genus (Maskulin, Feminin, Neutrum), serta konstruksi khas berupa trennbare Verben atau verba terpisah—fenomena yang tidak dijumpai dalam struktur bahasa Indonesia. Sebagai standar anotasi, digunakan Stuttgart-Tübingen-Tagset (STTS) yang digagas oleh Thielen et al. (1999) dan telah menjadi acuan utama dalam pengembangan korpus TIGER.

Data kajian bersumber dari teks naratif berjudul "Das Gewitter" yang ditulis oleh Ernst H. Gombrich sebagai bagian dari karya populernya, *Eine kurze Weltgeschichte für junge Leser* (1985/2002). Teks ini dipilih karena merepresentasikan genre narasi historis berbahasa Jerman dengan keragaman kelas kata yang kaya, sekaligus memuat rekaman peristiwa-peristiwa bersejarah signifikan—mulai dari keruntuhan Kekaisaran Romawi Barat, ekspansi bangsa Hunni, hingga penobatan Romulus Augustulus sebagai kaisar pamungkas Romawi pada tahun 476 Masehi.

Adapun tujuan penelitian ini mencakup tiga hal pokok: (1) mengidentifikasi dan mendeskripsikan sepuluh kategori kata (Wortarten) yang terdapat dalam teks tersebut; (2) menelaah pola distribusi frekuensi tiap kelas kata melalui metode komputasional; serta (3) mengungkap fenomena-fenomena linguistik yang berpotensi menjadi kendala dalam sistem POS tagging otomatis.

LANDASAN TEORI

1. Part-of-Speech Tagging

POS Tagging merupakan suatu proses anotasi yang memberikan label gramatikal dan sintaksis pada setiap kata dalam sebuah kalimat (Manning & Schütze, 1999). Sistem penandaan modern yang bertumpu pada pendekatan Hidden Markov Model (HMM), Maximum Entropy, maupun Conditional Random Field (CRF) mampu meraih tingkat akurasi di atas 97% untuk bahasa Inggris. Namun demikian, kinerja sistem serupa pada bahasa Jerman cenderung lebih rendah, yakni berkisar antara 93 hingga 96%, disebabkan oleh tingginya kompleksitas sistem morfologi yang dimiliki bahasa tersebut (Schmid, 1994).

2. Stuttgart-Tübingen-Tagset (STTS)

STTS merupakan standar tagset untuk bahasa Jerman yang terdiri dari 54 tag morfosintaksis (Thielen et al. , 1999). Tagset ini membedakan, contohnya, antara artikel (ART), pronomen demonstratif (PDS), dan pronomen kepemilikan (PPOSS). Kata kerja dibedakan menjadi kata kerja finit (VVFİN), infinitif (VVINF), dan partisip (VVPP). Tagset ini diterapkan dalam korpus TIGER (Brants et al. , 2004), yang memiliki lebih dari 900. 000 token teks berita dalam bahasa Jerman.

3. Wortarten dalam Bahasa Jerman

Tata bahasa Jerman yang klasik (Duden Grammatik, 2016) membagi kata-kata menjadi sepuluh kategori utama: Substantiv (Nomen), Verb, Adjektiv, Adverb, Artikel, Pronomen, Präposition, Konjunktion, Numerale, dan Partikel. Masing-masing kategori memiliki ciri morfologis, sintaksis, dan semantis yang membedakannya, walaupun batasan antara kategori-kategori tersebut tidak selalu jelas karena adanya konversi dan homonimi.

METODE PENELITIAN

Penelitian ini menerapkan pendekatan deskriptif yang memadukan metode kualitatif dan kuantitatif. Sumber data berasal dari teks "Das Gewitter" yang mengandung kurang lebih 300 kata, dengan sekitar 250 token yang dipilih secara purposif sebagai unit analisis. Prosedur penelitian dilaksanakan melalui tahapan-tahapan berikut:

1. Tokenisasi: teks diurai menjadi satuan-satuan token berdasarkan batas antarkata serta keberadaan tanda baca.
2. Anotasi manual: setiap token ditetapkan label kelas katanya dengan mengacu pada tagset

STTS dan panduan Duden Grammatik.

3. Verifikasi silang: dua anotator melakukan proses anotasi secara independen, dan setiap perbedaan hasil diselesaikan melalui mekanisme diskusi bersama.
4. Analisis distribusi frekuensi sekaligus identifikasi terhadap fenomena ambiguitas kategorial yang muncul dalam data.

Instrumen utama yang digunakan adalah tabel anotasi POS berbasis tagset STTS. Analisis data dijalankan secara deskriptif dengan menghitung frekuensi absolut maupun relatif dari masing-masing kelas kata, sekaligus menelaah pola distribusi yang mencerminkan register dan karakteristik teks yang dikaji.

HASIL DAN PEMBAHASAN

1. Distribusi Frekuensi 10 Jenis Kata

Hasil anotasi POS terhadap 250 token teks “Das Gewitter” menghasilkan distribusi frekuensi sebagaimana disajikan pada Tabel 1 berikut.

Tabel 1. Distribusi Frekuensi 10 Jenis Kata dalam Teks “Das Gewitter”

No.	Jenis Kata (Wortart)	Tag STTS	Frekuensi	Persentase	Fungsi Utama
1	Nomen (Substantiv)	NN / NE	68	27,2%	Referensi entitas
2	Verb	VV / VA / VM	52	20,8%	Predikat klausa
3	Artikel	ART	38	15,2%	Determinasi + Kasus
4	Pronomen	PPER / PIS / PRF	29	11,6%	Substitusi nomina
5	Präposition	APPR / APPRART	22	8,8%	Hubungan spasial/temporal

No.	Jenis Kata (Wortart)	Tag STTS	Frekuensi	Persentase	Fungsi Utama
6	Adjektiv	ADJA / ADJD	18	7,2%	Atribut / predikatif
7	Konjunktion	KON / KOUS	14	5,6%	Koordinasi & subordinasi
8	Adverb	ADV	12	4,8%	Modifikasi verba/adjektif
9	Numerale	CARD / ORD	11	4,4%	Referensi kronologis
10	Partikel	PTKNEG / PTKGT	6	2,4%	Negasi; gradasi
	Total		270	100%	

Data pada Tabel 1 menunjukkan bahwa Nomen merupakan kelas kata paling dominan dengan frekuensi 68 token (27,2%), diikuti Verb dengan 52 token (20,8%). Temuan ini sesuai dengan karakteristik teks naratif historis yang memiliki kepadatan referensial tinggi, yakni banyak mengacu pada entitas (tokoh, tempat, peristiwa) dan mendeskripsikan rangkaian tindakan (Biber, 1988). Artikel menempati posisi ketiga (15,2%), yang dapat dijelaskan oleh tata bahasa Jerman yang mewajibkan artikel pada hampir setiap nomina.

2. Analisis Per Kelas Kata

a. Nomen (Substantiv)

Nomen pada bahasa Jerman pastinya selalu ditulis dengan menggunakan huruf kapital, yang secara komputasional memudahkan identifikasi awal tetapi mengakibatkan ambiguitas dengan kata-kata lain yang dikapitalkan (contohnya mulai dari awal kalimat, Eigennamen). Dari 68 token Nomen, 12 di antaranya adalah Eigennamen (nama diri) contohnya seperti Attila, Gallien, Konstantinopel, dan Romulus Augustulus. Dalam terminologi STTS, Appellativa diberi tag NN sedangkan Eigennamen diberi tag NE.

Contoh dalam teks: das Gewitter, die Hunnen, der Kaiser, das Reich, das Gewitter, die

Berge.

b. Verb (Zeitwort)

Verba pada teks memakai berbagai bentuk morfologis. Teks ini didominasi oleh Präteritum (bentuk lampau) yang merupakan khas untuk narasi historis (Weinrich, 1964), seperti kamen, hatte, hieß, wurde. Terdapat pula Verba bantu (Hilfsverben): haben, sein, dan werden untuk konstruksi Perfekt, Plusquamperfekt, dan Passiv. Trennbare Verben seperti steigen auf dan aufziehen membutuhkan penanganan khusus dalam tagging dikarenakan prefiks verba berpindah ke akhir klausa.

Contohnya: sieht, fühlt, war, kamen, wurde abgesetzt, steigen...auf, aufziehen.

c. Artikel

Artikel dalam bahasa Jerman berfungsi ganda: menunjukkan definiteness (Bestimmter Artikel: der/die/das) atau indefiniteness (Unbestimmter Artikel: ein/eine) sekaligus juga mengkodekan Kasus dan Genus. Dalam teks diketahui bahwa Verschmelzungsform (fusi preposisi dan artikel), yaitu im (in+dem), am, zum, dan zur. Fenomena ini merupakan suatu tantangan tokenisasi dalam NLP karena satu ortografis menggambarkan dua morfem gramatikal.

d. Pronomen

Pronomen dalam teks meliputi lima subkelas: (1) Personalpronomen (er, sie, es, du, man), (2) Reflexivpronomen (sich), (3) Indefinitpronomen (man, nichts, etwas), (4) Relativpronomen (die, der), dan (5) Demonstrativpronomen (das). Yang merupakan sangat menarik secara komputasional adalah pronomen “man” yang ketika pronomen generik tanpa padanan langsung dalam bahasa Indonesia, sering diterjemahkan sebagai ‘seseorang’ atau konstruksi pasif.

e. Präposition

Präposition mengharuskan kasus tertentu pada nomina/pronomen yang mematuhinya. Teks memperlihatkan präposition rektional kasus tetap (mit +Dat, von +Dat, nach +Dat, gegen +Akk) dan präposition dengan kasus saling berganti (Wechselpräpositionen: in, an, auf yang digunakan dengan Dativ untuk posisi atau Akkusativ untuk perubahan). Distribusi präposition (8,8%) tetap stabil dengan penelitian terhadap teks naratif bahasa Jerman (Meurers, 2012).

f. Adjektiv

Adjektiv yang berfungsi sebagai atributif (ADJA) mengalami perubahan bentuk sesuai dengan artikel yang mendahuluinya (deklinsi lemah/kuat/campuran), sedangkan adjektiv yang berperan sebagai predikatif (ADJD) tidak mengalami deklinsi. Selain itu, terdapat Partizip II yang berfungsi sebagai adjektiv atributif, contohnya geballte Wolken. Bentuk perbandingan ditemukan dalam bentuk Komparativ: näher (dekat → lebih dekat).

g. Konjunktion

Konjungsi koordinatif (KON) dalam tulisan meliputi und (kopulatif), aber (adversatif), und doch (konsesif), serta denn (kausal). Sementara itu, konjungsi subordinatif (KOUS) mencakup dass (objektif), wenn (kondisional), als (temporal), dan obwohl (konsesif). Dari segi sintaksis, KOUS menyebabkan pergeseran verba ke posisi akhir (Verbletzt-Stellung) dalam klausa subordinat, yang menjadi indikator penting klausa subordinat dalam konteks komputasi.

h. Adverb

Adverb tidak mengalami perubahan bentuk (unveränderlich) sehingga relatif mudah untuk dikenali. Adverb temporal (schon, dann, erst, endlich) paling banyak muncul dalam teks, mencerminkan struktur naratif yang menekankan urutan waktu. Adverb lokal (dort, da, weg) dan adverb modal (besonders, recht, wirklich) juga dapat ditemukan. Tantangan muncul dengan adverb polisemi seperti da yang dapat berfungsi sebagai adverb lokal ('di sana') atau konjunktur kausal ('karena').

i. Numerale (Zahlwort)

Numerale dalam teks didominasi oleh Kardinalzahl untuk merujuk pada tahun-tahun sejarah: 444, 410, 451, 453, dan 476 (semuanya "nach Christus" atau "nach Christi Geburt"). Pemanfaatan angka sebagai penanda waktu adalah ciri utama dalam teks sejarah. Token dapat bersifat ambigu: dapat dipahami sebagai Numerale ('satu kali') atau Adverb ('suatu ketika'), tergantung pada konteks.

j. Partikel

Meskipun memiliki frekuensi terendah (2,4%), partikel memiliki manfaat praktis yang signifikan. Nilai kebenaran proposisi dibalik oleh partikel negation nicht dan kein. Negasi diperkuat oleh gradpartikel gar (dalam gar nichts). Abtönungspartikel ja, doch, und doch membangun kohesi wacana dan memperhalus modalitas penutur. Karena maknanya sangat

bergantung pada konteks, Abtönungspartikel termasuk kategori yang paling sulit untuk diklasifikasikan secara otomatis dalam pemrosesan bahasa alami.

k. Fenomena Linguistik Komputasional

Dari analisis tersebut, teridentifikasi enam fenomena yang penting untuk pengembangan sistem POS tagging bahasa Jerman yang dirangkum dalam Tabel 2.

Tabel 2. Fenomena Linguistik Komputasional dalam Teks

Fenomena	Contoh dari Teks	Tantangan Komputasional
Ambiguitas Artikel/Pronomen	<i>das Gewitter</i> vs. <i>Das ist großartig</i>	ART vs. PDS — perlu analisis konteks sintaksis
Trennbare Verben	<i>steigen ... auf, aufziehen</i>	Prefiks terpisah jauh dari verba; perlu co-reference resolution
Verschmelzungsform	<i>im (in+dem), am, zum</i>	Satu token = dua morfem; perlu morfem-splitting
Konversion (V→N, V→ADJ)	<i>das Donnern (N), geballte Wolken (ADJ)</i>	Nominalisasi infinitif; Partizip sebagai ADJA
Ambiguitas ADV/KON	<i>dann, doch, aber</i>	Posisi klausa menentukan tag; perlu analisis kalimat penuh
Eigennamen (NE)	<i>Attila, Gallien, Romulus Augustulus</i>	Kapital mirip NN; perlu NER dan gazetteer

Fenomena trennbare Verben merupakan tantangan terbesar dalam tagging otomatis bahasa Jerman. Algoritme berbasis window lokal sering gagal menghubungkan prefiks verba yang terpisah lebih dari 10 token dari verba induknya (Uszkoreit, 1987). Pendekatan berbasis dependensi (dependency parsing) lebih efektif untuk menangani fenomena ini.

KESIMPULAN

Penelitian ini berhasil mengenali dan menganalisis 10 tipe kata dalam teks naratif sejarah “Das Gewitter” oleh Ernst H. Gombrich dengan menggunakan metode komputasi linguistik

berdasar STTS. Kesimpulan utama dari penelitian ini adalah:

1. Nomen adalah kelas kata yang paling dominan (27,2%) diikuti oleh Verb (20,8%) dan Artikel (15,2%), mencerminkan ciri-ciri dari register naratif historis dengan kepadatan referensial dan temporal yang tinggi.
2. Numerale berfungsi sebagai penanda waktu dalam sejarah, sedangkan Partikel memiliki peran kohesi wacana yang penting meskipun jumlahnya sedikit.
3. Terdapat enam fenomena linguistik yang diidentifikasi sebagai tantangan utama untuk sistem otomatis POS tagging: ambiguitas ART/PDS, trennbare Verben, bentuk penggabungan, konversi kelas kata, ambiguitas ADV/KON, dan nama diri.

Studi ini diharapkan mampu memberikan sumbangan bagi pengembangan sumber daya bahasa Jerman (terutama anotasi POS dalam genre sejarah populer) dan menjadi acuan bagi pengajaran bahasa Jerman yang berbasis korpus. Penelitian lebih lanjut dianjurkan untuk memanfaatkan parser otomatis (seperti TreeTagger atau spaCy de) guna memverifikasi hasil anotasi manual serta menganalisis korpus yang lebih luas.

DAFTAR PUSTAKA

- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.
- Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, G., Lezius, W., Schulte im Walde, S., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4), 597–620.
- Duden. (2016). *Duden Grammatik der deutschen Gegenwartssprache* (9. Auflage). Bibliographisches Institut.
- Gombrich, E. H. (2002). *Eine kurze Weltgeschichte für junge Leser*. DuMont Buchverlag. (Original: 1985)
- Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed., Draft). <https://web.stanford.edu/~jurafsky/slp3/>
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Meurers, W. D. (2012). *Natural Language Processing and Language Learning*. In *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*.

-
- Thielen, C., Schiller, A., Teufel, S., & Stöckert, C. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. Universität Stuttgart / Universität Tübingen.
- Uszkoreit, H. (1987). Word Order and Constituent Structure in German. CSLI Publications.
- Weinrich, H. (1964). Tempus: Besprochene und erzählte Welt. Kohlhammer.