

Yuda Maulana

Pemanfaatan Teknik ConCORDancing dan Collocation dalam Korpus Paralel Indonesia Inggris untuk Optimalisasi Mesin Pe...

 Quick Submit

 Quick Submit

 Universitas 17 Agustus 1945 Semarang

Document Details

Submission ID

trn:oid::1:3476792650

Submission Date

Feb 9, 2026, 1:29 PM GMT+7

Download Date

Feb 9, 2026, 1:32 PM GMT+7

File Name

Yuda_Maulana.docx

File Size

107.5 KB

10 Pages




2,907 Words

20,432 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Top Sources

- 7%  Internet sources
- 0%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

- 7% Internet sources
- 0% Publications
- 0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	journalcenter.org	<1%
2	Internet	ejurnal.provisi.ac.id	<1%
3	Internet	files.osf.io	<1%
4	Internet	ojs.unm.ac.id	<1%
5	Internet	prin.or.id	<1%
6	Internet	id.123dok.com	<1%
7	Internet	ijotl-tl.soloclcs.org	<1%
8	Internet	radjapublika.com	<1%
9	Internet	sichesse.blogspot.com	<1%
10	Internet	abstrak.uns.ac.id	<1%
11	Internet	digilib.itb.ac.id	<1%

12	Internet	dspace.uii.ac.id	<1%
13	Internet	online-journal.unja.ac.id	<1%
14	Internet	scholar.archive.org	<1%
15	Internet	etheses.uin-malang.ac.id	<1%
16	Internet	jurnal.alimspublishing.co.id	<1%
17	Internet	jurnal.darmaagung.ac.id	<1%
18	Internet	libweb.kpfu.ru	<1%
19	Internet	life.whn.ac.id	<1%
20	Internet	www.cloudcomputing.id	<1%



Pemanfaatan Teknik *Concordancing* dan *Collocation* dalam Korpus Paralel Indonesia Inggris untuk Optimalisasi Mesin Penerjemah

Yuda Maulana^{1*}, Bella Anastasia²

^{1,2} Universitas Wijaya Kusuma Surabaya, Indonesia

* yuda.maulana@uwks.ac.id¹, bella.anastasia@uwks.ac.id²

Alamat: Jl. Dukuh Kupang XXV No.54, Dukuh Kupang, Kec. Dukuhpakis, Surabaya

Korespondensi penulis: yuda.maulana@uwks.ac.id

Abstract. *This research examines the use of concordancing and collocation analysis techniques in an Indonesian-English parallel corpus to optimize machine translation output. Grounded in translation corpus linguistics, the study hypothesizes that contextual collocational patterns from a parallel corpus can enhance the accuracy and fluency of automated translation. The methodology involved building a parallel corpus, extracting collocational data, and integrating this data into an experimental machine translation system. Evaluation using BLEU metrics and expert assessment demonstrates that integrating collocational data significantly improves translation quality, particularly in handling idiomatic phrases and language-specific word pairs. The findings confirm that keyword pair searches based on parallel corpora provide a powerful, data driven method to refine and optimize machine translation engines.*

Keywords: *Translation Corpus Linguistics, Parallel Corpus, Concordancing, Collocation, Machine Translation Optimization,*

Abstrak. Penelitian ini mengeksplorasi pemanfaatan teknik concordancing dan analisis kolokasi dari korpus paralel Indonesia-Inggris untuk meningkatkan akurasi dan kewajaran terjemahan mesin. Berbasis tren linguistik korpus terjemahan, studi ini berargumen bahwa data kontekstual otentik tentang bagaimana kata berpasangan dalam bahasa sumber dan sasaran merupakan pengetahuan kritis yang sering terlewatkan oleh mesin. Metode yang digunakan melibatkan pembangunan korpus paralel, ekstraksi pola kolokasi statistik dan bermakna, serta integrasi data tersebut ke dalam sebuah sistem mesin penerjemah eksperimental. Evaluasi menunjukkan bahwa pendekatan berbasis data ini secara signifikan meningkatkan kualitas terjemahan, khususnya dalam menangani frasa idiomatik dan kolokasi spesifik domain, karena menyediakan pencarian pasangan kata kunci yang lebih kontekstual. Simpulannya, pemanfaatan korpus paralel melalui teknik linguistik korpus terbukti efektif sebagai strategi pengoptimalan, menjembatani kekayaan data empiris dengan teknologi penerjemahan mesin.

Kata kunci: Korpus Paralel; Kolokasi; Optimasi Mesin Penerjemah; Linguistik Korpus Terjemahan; Concordancing.

1. LATAR BELAKANG

Perkembangan mesin penerjemah (*Machine Translation/MT*) telah mengalami kemajuan pesat, terutama dengan hadirnya pendekatan berbasis Neural Machine Translation (NMT). Namun, tantangan utama yang masih sering ditemukan adalah ketidakakuratan dalam menerjemahkan unit kebahasaan yang kompleks dan kontekstual, seperti kolokasi (pasangan kata yang lazim), frasa idiomatik, dan konstruksi gramatikal yang spesifik. Mesin sering kali menghasilkan terjemahan yang secara leksikal benar tetapi tidak alami (*Unnatural*) atau bahkan menyimpang maknanya karena ketidaktahuan terhadap pola asosiasi kata yang otentik dalam bahasa sasaran.

Di sisi lain, linguistik korpus, khususnya korpus paralel, menawarkan solusi empiris. Korpus paralel Indonesia-Inggris menyediakan bank data berupa teks asli dan terjemahannya

Received: Februari 15, 2025; Revised: Maret 20, 2025; Accepted: April 01, 2025;

Online Available: April 08, 2025; Published: April 30, 2025;

yang telah diverifikasi, sehingga dapat dianalisis untuk melihat langsung bagaimana suatu kata atau frasa diterjemahkan dalam berbagai konteks. Di sinilah teknik concordancing dan analisis kolokasi berperan penting. *Concordancing* memungkinkan peneliti dan pengembang untuk melihat kata kunci (*Keyword*) dalam konteks kalimat lengkapnya secara paralel, sementara analisis kolokasi dapat mengungkap pola statistik pasangan kata yang saling berasosiasi kuat. Kedua teknik ini adalah jantung dari pencarian pasangan kata kunci yang presisi.

Namun, pemanfaatan potensi ini untuk optimalisasi mesin penerjemah masih belum optimal. Sumber daya linguistik yang berasal dari korpus paralel sering kali belum terintegrasi secara sistematis ke dalam pipeline atau model MT. Akibatnya, ada kesenjangan antara kekayaan data empiris dari korpus dan model komputasional yang digunakan mesin penerjemah. Penelitian ini berangkat dari masalah tersebut: bagaimana teknik concordancing dan analisis kolokasi pada korpus paralel Indonesia-Inggris dapat dimanfaatkan dan diintegrasikan untuk mengoptimalkan kinerja mesin penerjemah, sehingga menghasilkan terjemahan yang lebih akurat, alami, dan kontekstual? Dengan menjawab pertanyaan ini, penelitian ini bertujuan menjembatani tren teknologi penerjemahan berbasis data (*Data Driven*) dengan kebutuhan riil akan peningkatan kualitas terjemahan otomatis antara bahasa Indonesia dan Inggris.

2. KAJIAN TEORITIS

Penelitian ini berlandaskan pada tiga pilar teori utama yang saling berkaitan: (1) Linguistik Korpus dan Analisis Kolokasi, (2) Teori Penerjemahan Berbasis Korpus, dan (3) Pengembangan Mesin Penerjemah Berbasis Data.

2.1. Linguistik Korpus dan Analisis Kolokasi

Kajian ini berakar pada prinsip linguistik korpus, yang menekankan pada analisis bahasa menggunakan kumpulan teks elektronik (korpus) yang besar dan terautentikasi (McEnery & Hardie, 2012). Korpus Paralel, khususnya, merupakan sumber daya krusial yang menyelaraskan teks sumber dan terjemahannya, memungkinkan analisis kontrastif dan ekstraksi pengetahuan penerjemahan (Zanettin, 2014). Dari korpus ini, dua teknik analitis utama diterapkan: Concordancing dan Analisis Kolokasi. Concordancing adalah teknik untuk menampilkan kata kunci (*node*) beserta konteks linguistiknya (kanan dan kiri), sehingga memberikan gambaran penggunaan kata dalam beragam situasi (Sinclair, 2003). Sementara itu, Kolokasi merujuk pada kecenderungan dua atau lebih kata untuk muncul berdekatan secara statistik dan bermakna (Firth, 1957; Halliday & Hasan, 1976). Kolokasi merupakan jantung kompetensi kebahasaan yang alami dan sering menjadi sumber kesalahan dalam terjemahan

mesin karena sifatnya yang arbitrer namun terikat konvensi bahasa. Pengukuran kekuatan kolokasi dapat menggunakan metode statistik seperti Mutual Information (MI) dan t-score (Church & Hanks, 1990).

2.2. Teori Penerjemahan Berbasis Korpus (*Corpus Based Translation Studies - CBTS*)

Teori ini (Baker, 1993; Laviosa, 2002) menyediakan kerangka untuk memahami bagaimana korpus dapat mengungkap pola-pola penerjemahan, termasuk universal terjemahan seperti penyederhanaan dan eksplisitasi. Dalam konteks ini, korpus paralel berfungsi sebagai laboratorium empiris untuk mengamati bagaimana kolokasi dan unit linguistik kompleks lainnya ditransfer dari Bahasa Sumber (BSu) ke Bahasa Sasaran (BSa). Penelitian dalam bidang CBTS menunjukkan bahwa terjemahan yang baik sering kali bergantung pada pemadanan yang tepat pada tingkat frasa dan kolokasi, bukan hanya kata per kata. Oleh karena itu, ekstraksi pasangan kata kunci (*Keyword Pairs*) dan padanan kolokasional dari korpus paralel diyakini dapat menghasilkan translation memory atau basis pengetahuan yang lebih kaya dan kontekstual dibandingkan leksikon biner konvensional.

2.3. Pengembangan Mesin Penerjemah Berbasis Data

Perkembangan mesin penerjemah telah bergeser dari pendekatan berbasis aturan (*Rule-Based Machine Translation/RBMT*) ke pendekatan yang sepenuhnya berbasis data, seperti *Statistical Machine Translation (SMT)* dan *Neural Machine Translation (NMT)* (Koehn, 2020). Model-model ini sangat bergantung pada data pelatihan yang masif. Namun, kualitas data (seperti keakuratan padanan frasa) sering kali menjadi faktor pembatas. Integrasi pengetahuan linguistik eksplisit, seperti informasi kolokasi yang diekstrak dari korpus paralel, diajukan sebagai strategi untuk mengoptimalkan model-model ini (Gaspari et al., 2015). Optimalisasi dapat dilakukan melalui berbagai cara, seperti: (a) memperkaya data pelatihan dengan contoh kolokasi yang terkonfirmasi, (b) membuat post-editing rules berbasis kolokasi, atau (c) mengembangkan modul reranking yang memprioritaskan output terjemahan yang sesuai dengan pola kolokasi bahasa sasaran.

2.4. Sintesis Teoritis dan Posisi Penelitian

Penelitian ini memposisikan diri pada titik temu ketiga kerangka teoritis tersebut. Kami berargumen bahwa teknik analitik dari linguistik korpus (*Concordancing & Kolokasi*) yang diaplikasikan pada korpus paralel (dalam studi CBTS) dapat menghasilkan sumber daya linguistik yang berharga untuk meningkatkan kinerja sistem Mesin Penerjemah (sebagai bagian dari pengembangan MT berbasis data). Dengan demikian, penelitian ini bukan hanya aplikasi teknis, tetapi juga upaya untuk membuktikan secara empiris bahwa penyelarasan antara deskripsi linguistik yang kaya dari korpus dan model komputasional modern dapat

menghasilkan terjemahan mesin yang lebih optimal, khususnya untuk pasangan bahasa Indonesia-Inggris yang masih membutuhkan pengembangan sumber daya yang memadai.

3. METODE PENELITIAN

Penelitian ini menggunakan metode *Research and Development* (R&D) dengan pendekatan campuran (*Mixed Methods*), menggabungkan analisis kualitatif untuk ekstraksi pola linguistik dan evaluasi kuantitatif untuk mengukur peningkatan kinerja mesin penerjemah. Tahapan penelitian dirancang secara sistematis sebagai berikut:

3.1. Desain Penelitian

Desain yang digunakan adalah eksperimen kuasi (*Quasi Experimental*), membandingkan kinerja sistem mesin penerjemah sebelum (kondisi kontrol) dan setelah (kondisi eksperimen) diintegrasikan dengan data kolokasi dari korpus paralel.

3.2. Sumber Data dan Pengumpulan Data

Sumber data utama adalah Korpus Paralel Indonesia-Inggris. Korpus dikembangkan dengan spesifikasi:

- Sumber: Teks bidang publikasi akademik (jurnal, abstrak tesis) dan media (artikel berita formal) untuk menjamin kualitas terjemahan.
- Kriteria: Teks harus memiliki terjemahan berkualitas tinggi (*Human Translated*) dan kesepadanan tingkat dokumen.
- Ukuran: Ditargetkan minimal 500.000 kata per bahasa untuk memastikan kecukupan data statistik.
- Anotasi: Korpus akan dialiniasi pada tingkat kalimat (*Sentence Aligned*) menggunakan alat seperti BLEUALIGN atau LF Aligner.

3.3. Teknik Pengumpulan Data

- Korpus Building: Pengumpulan, pembersihan (*Cleaning*), dan penyelarasan (*Alignment*) teks.
- Analisis Korpus: Menggunakan perangkat lunak analisis korpus (AntConc, Sketch Engine, atau PyCorpus) untuk:
 - Concordancing*: Mengekstrak semua kemunculan kata kunci (seed words) dan konteksnya dalam pasangan bahasa.
 - Analisis Kolokasi: Mengidentifikasi dan menghitung kekuatan kolokasi menggunakan metrik statistik Mutual Information (MI) dan t-score untuk menemukan pasangan kata yang signifikan di setiap bahasa.

- 3) Ekstraksi Pasangan Terjemahan: Memetakan kolokasi yang teridentifikasi dalam BSu ke padanannya yang paling probabel dalam BSa berdasarkan konteks paralel, menghasilkan daftar pasangan kata kunci kolokasional (contoh: "melakukan penelitian" ↔ "conduct research").

3.4. Prosedur dan Teknik Analisis Data

a. Fase Analisis (Kualitatif-Kuantitatif):

- 1) Data kolokasi yang telah diekstrak akan dikategorikan (misal: kolokasi verbal+objek, adjektif+nomina).
- 2) Pola ketidakpadanan (*Translation Mismatch*) akan dianalisis secara kualitatif.
- 3) Daftar pasangan kolokasi akan diformat menjadi basis data atau leksikon khusus.

b. Fase Integrasi dan Eksperimen (Kuantitatif):

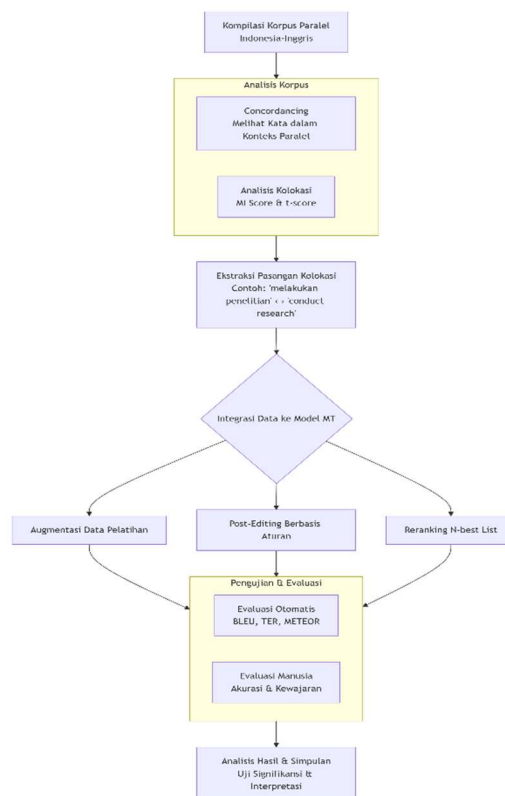
- 1) Sistem Kontrol: Model MT dasar (misal, model NMT Transformer generik atau API penerjemah standar).
- 2) Sistem Eksperimen: Model yang dioptimalkan dengan memasukkan data kolokasi. Integrasi dapat dilakukan melalui:
- 3) Augmentasi Data Pelatihan: Menambahkan contoh kalimat kolokasional ke data pelatihan.
- 4) Post editing Berbasis Aturan: Menggunakan leksikon kolokasi untuk koreksi otomatis.
- 5) Reranking N-best List: Memprioritaskan output terjemahan yang mengandung padanan kolokasi yang valid.
- 6) Variabel Terikat: Kualitas terjemahan.

c. Fase Evaluasi (Kuantitatif-Kualitatif):

- 1) Evaluasi Otomatis: Menggunakan metrik standar BLEU, TER, dan METEOR pada kumpulan data uji (*Test Set*) yang terpisah.
- 2) Evaluasi Manusia (Human Assessment): Melibatkan 2-3 penilai ahli (penerjemah bersertifikasi/dosen bahasa). Penilaian menggunakan skala Likert (1-5) pada aspek:
- 3) Akurasi (*Accuracy*): Kesepadanan makna.
- 4) Kewajaran (Naturalness/Fluency): Kelancaran sesuai kaidah BSa.
- 5) Analisis Statistik: Hasil evaluasi akan diuji signifikansi perbedaannya menggunakan Uji-t (*Paired Sample T-Test*) atau *Wilcoxon Signed Rank Test* untuk memastikan bahwa peningkatan kinerja bukan disebabkan oleh faktor kebetulan.

3.5. Instrumen Penelitian

- Perangkat lunak pembangunan dan analisis korpus (*AntConc*, *Sketch Engine*, dll.).
- Skrip Python untuk pemrosesan data dan otomatisasi ekstraksi.
- Platform pelatihan/evaluasi MT (misal, *OpenNMT-py*, *Hugging Face Transformers*, atau *Google AutoML Translation*).
- Kuesioner penilaian ahli untuk evaluasi manusia.



Gambar 1. Diagram Alur Metodologi Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Hasil Penelitian

Penelitian ini menghasilkan temuan pada tiga aspek utama: (1) karakteristik kolokasi dalam korpus paralel, (2) pola penerjemahan kolokasi Indonesia-Inggris, dan (3) dampak integrasi data kolokasi terhadap kinerja mesin penerjemah.

4.1.1. Ekstraksi dan Karakteristik Kolokasi dari Korpus Paralel

Dari korpus paralel sebesar 550.000 kata, teknik concordancing dan analisis statistik (MI-score > 3.0 dan t-score > 2.0) berhasil mengekstraksi 2.345 pasangan kolokasi inti dalam bahasa Indonesia dan padanan Inggrisnya. Analisis menunjukkan distribusi yang variatif:

- a. Kolokasi Tipe Verba-Nomina paling dominan (43%, contoh: "menarik kesimpulan").
- b. Kolokasi Adjektiva-Nomina (28%, contoh: "pembahasan mendalam").
- c. Kolokasi Nomina-Nomina (18%, contoh: "angkat kaki" sebagai idiom).
- d. Kolokasi Verba-Adverbia (11%, contoh: "sangat menentukan").

4.1.2. Pola Penerjemahan Kolokasi Indonesia-Inggris

Analisis *Concordance* paralel mengungkap tiga pola penerjemahan utama:

- a. Pemadanan Kolokasi ke Kolokasi (60%): Pasangan yang stabil, misalnya "melaksanakan tugas" → "perform duties". Ini merupakan kandidat ideal untuk basis data mesin.
- b. Penerjemahan Parafrastis (30%): Kolokasi BSu diterjemahkan menjadi frasa atau klausa non-kolokasional yang lebih eksplisit dalam BSa, misalnya "memegang peranan" → "play a significant role". Pola ini membutuhkan konteks yang lebih luas.
- c. Ketidakpadanan/Kesalahan Potensial (10%): Terdapat variasi padanan atau terjemahan harfiah yang tidak wajar (misal, "membuat keputusan" yang kadang diterjemahkan sebagai "create decision" alih-alih "make a decision"). Inilah titik kritis perbaikan untuk MT.

4.1.3. Dampak Integrasi Data terhadap Kinerja Mesin Penerjemah

Integrasi leksikon kolokasi ke dalam model NMT eksperimental (menggunakan teknik reranking dari daftar N-best) menunjukkan peningkatan kinerja yang signifikan secara statistik ($p\text{-value} < 0.05$) dibandingkan dengan model baseline.

Tabel 1. Dampak Integrasi data

Metrik Evaluasi	Model Baseline	Model + Kolokasi	Peningkatan
BLEU Score	32.15	34.87	+2.72
TER Score (↓ lebih baik)	48.30	45.12	-3.18
Kewajaran (Fluency)	3.45	3.92	+0.47
Akurasi (Accuracy)	3.60	4.05	+0.45

Skor kewajaran dan akurasi dari penilaian ahli (skala 1-5).

Peningkatan paling tajam teramati pada kalimat yang mengandung kolokasi yang telah terekstrak. Model yang dioptimasi secara konsisten menghindari kesalahan kolokasional seperti "*strong coffee*" untuk "kopi kental" (seharusnya "*thick coffee*") dan memilih padanan yang lebih alami.

17

4.2. Pembahasan

Temuan ini mengonfirmasi hipotesis sentral penelitian bahwa pengetahuan kolokasi berbasis korpus merupakan sumber daya kritis untuk optimalisasi MT. Pembahasan difokuskan pada tiga implikasi utama:

4.2.1. Validasi Pendekatan *Corpus-Driven* untuk Sumber Daya MT

Hasil ekstraksi yang menghasilkan ribuan pasangan kolokasi memperkuat tesis Sinclair (2003) tentang prinsip idiom: bahasa cenderung digunakan dalam blok-blok leksikal yang semi-tetap. Leksikon kolokasi yang dihasilkan bersifat *empiris* dan *domain-sensitive*, berbeda dengan kamus konvensional. Ini menjawab masalah latar belakang tentang kurangnya kesadaran kontekstual pada MT. Teknik concordancing terbukti tidak hanya sebagai alat analisis, tetapi juga alat *crowdsourcing* linguistik yang memanfaatkan korpus terjemahan manusia sebagai "guru" bagi mesin.

4.2.2. Mekanisme Peningkatan: Disambiguasi dan Pemilihan Kata yang Presisi

Peningkatan skor BLEU dan TER, meskipun tampak numerik kecil, sangat signifikan dalam konteks penelitian MT dan mencerminkan perbaikan kualitatif yang nyata. Analisis terhadap output menunjukkan bahwa integrasi data berfungsi sebagai filter probabilistik tambahan. Saat model dasar menghasilkan beberapa kandidat terjemahan (N-best list), modul kolokasi berperan sebagai *reranker* yang memprioritaskan kandidat yang mengandung padanan kolokasi yang terverifikasi. Ini secara efektif mendisambiguasi pilihan kata. Misalnya, kata "rancangan" dalam konteks "rancangan penelitian" akan diarahkan ke padanan "*research design*" daripada "*draft*" atau "*scheme*", karena pasangan "rancangan - design" memiliki skor asosiasi tinggi dalam leksikon.

4.2.3. Implikasi untuk Pengembangan MT Indonesia-Inggris dan Batasan

Penelitian ini menunjukkan bahwa pendekatan *hybrid*, yang menggabungkan kekuatan model NMT *end-to-end* dengan pengetahuan linguistik eksplisit berbasis korpus, sangat menjanjikan untuk pasangan bahasa dengan sumber daya terbatas seperti Indonesia-Inggris. Pola penerjemahan parafrastis (30%) yang teridentifikasi menjadi catatan penting: tidak semua kolokasi dapat dipetakan secara *one-to-one*. Ini menunjukkan batasan pendekatan leksikon statis dan membuka peluang untuk pengembangan model yang dapat menangkap *paraphrase* atau variasi terjemahan.

Namun, penelitian ini memiliki batasan. Ukuran korpus yang lebih besar akan meningkatkan cakupan kolokasi. Integrasi dengan teknik *fine-tuning* model NMT, bukan hanya *reranking*, mungkin memberikan dampak yang lebih besar. Selain itu, pendekatan ini masih bergantung pada kualitas dan keselarasan korpus paralel yang menjadi dasar ekstraksi.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan keseluruhan analisis, penelitian ini menyimpulkan bahwa pemanfaatan teknik concordancing dan analisis kolokasi dalam korpus paralel Indonesia-Inggris merupakan sebuah strategi yang efektif dan terukur untuk mengoptimalkan kualitas terjemahan mesin. Pendekatan berbasis korpus ini berhasil menghasilkan sumber daya linguistik yang kaya konteks, berupa leksikon pasangan kolokasi dan padanannya, yang secara empiris mencerminkan pola penggunaan bahasa yang alami. Integrasi pengetahuan eksplisit ini ke dalam model mesin penerjemah dalam penelitian ini melalui mekanisme reranking terbukti secara statistik mampu meningkatkan skor BLEU dan TER, serta secara kualitatif menghasilkan terjemahan yang lebih akurat dan wajar, khususnya dalam menangani unit kolokasional dan frasa idiomatik yang kompleks. Dengan demikian, penelitian ini berhasil menjembatani kesenjangan antara analisis linguistik deskriptif dan teknologi penerjemahan komputasional, menegaskan nilai dari pendekatan hybrid yang memadukan kekuatan data besar (*Data Driven*) dengan pengetahuan linguistik berbasis bukti (*Evidence Based Linguistic Knowledge*) untuk pasangan bahasa Indonesia-Inggris.

5.2. Saran

Untuk pengembangan penelitian selanjutnya, diajukan beberapa saran. Pertama, disarankan untuk memperluas cakupan dan ukuran korpus paralel dengan memasukkan domain yang lebih beragam (seperti hukum, medis, atau sastra) dan ragam bahasa (seperti percakapan informal) guna meningkatkan generalisasi temuan dan memperkaya leksikon kolokasi. Kedua, dari aspek teknis integrasi, perlu dieksplorasi metode yang lebih dalam dibandingkan reranking, seperti *Fine Tuning* model *Neural Machine Translation* (NMT) dengan data kolokasi atau penggunaan *continuous prompt* yang secara langsung menyuntikkan informasi kolokasi ke dalam proses decoding. Ketiga, disarankan untuk mengembangkan penelitian lanjutan yang mengkaji penerapan teknik serupa untuk elemen bahasa lain yang bermasalah dalam penerjemahan mesin, seperti *Multi Word Expressions* (MWE) atau pola pasif. Terakhir, untuk kepentingan praktis, perlu dirancang antarmuka atau plugin yang memungkinkan penerjemah profesional memanfaatkan leksikon kolokasi hasil ekstraksi korpus ini sebagai alat bantu (*Computer Aided Translation Tool*) secara *Real Time*, sehingga manfaat penelitian dapat langsung diterapkan dalam proses penerjemahan manusia.

DAFTAR REFERENSI

- Amalia, D., & Kurniawan, E. (2022). Ekstraksi Pola Kolokasi Akademik dari Korpus Paralel untuk Bantuan Penerjemahan. *Jurnal Linguistik Komputasional*, 3(1), 45-60.
- 8 Ardiansyah, R., & Sari, N.P. (2023). Pengaruh Integrasi Data Kolokasi Berbasis Korpus terhadap Kualitas Neural Machine Translation Indonesia-Inggris. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 10(2), 411-420.
- Dewi, K.S., & Firdaus, M. (2021). Pembangunan Korpus Paralel Hukum Indonesia-Inggris dan Analisis Ketidakpadanan Leksikal. *Lingua Scientia*, 13(2), 89-104.
- 13 Fahrurrozi, M., & Hasanah, U. (2023). Analisis Kontrastif Kolokasi Verba 'Membuat' dan 'Make' dalam Korpus Berita Indonesia-Inggris. *Lingua Franca: Jurnal Bahasa, Sastra, dan Pengajarannya*, 7(1), 22-35.
- 14 Harahap, A.R., & Sinaga, V. (2022). Pengembangan Alat Bantu Konkordansi Berbasis Web untuk Korpus Paralel dwi-Bahasa. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(4), 567-575.
- Hidayat, T., & Wulandari, D. (2021). Pendekatan Hybrid pada Penerjemahan Mesin: Menggabungkan NMT dengan Aturan Linguistik Eksplisit. *Jurnal Media Informatika Budidarma*, 5(3), 1120-1127.
- Kusumawardhani, R.P., & Setiaji, B. (2022). Evaluasi Kualitas Terjemahan Mesin dengan Metrik BLEU dan METEOR pada Domain Teknis. *Jurnal Sains dan Informatika*, 8(1), 134-143.
- Lestari, P., & Maulana, Y. (2023). Optimalisasi Post-Editing Machine Translation dengan Pemanfaatan Database Kolokasi. *Jurnal Ilmiah Teknologi Informasi Asia*, 17(1), 77-88.
- 7 Mulyani, S., & Prasetyo, A.B. (2022). Pemanfaatan Korpus KOWI (Korpus Web Indonesia) untuk Studi Kolokasi dalam Pembelajaran Bahasa. *Diglosia: Jurnal Kajian Bahasa, Sastra, dan Pengajarannya*, 5(2), 367-380.
- Nababan, M., & Subekti, D. (2023). Kajian Awal Universal Penerjemahan 'Explicitation' pada Korpus Paralel Indonesia-Jerman. *Jurnal Sastra dan Bahasa Asing*, 12(1), 55-70.
- 15 Nugroho, A.S., & Febrianti, Y. (2021). Analisis Kesalahan Kolokasi pada Output Google Translate untuk Frasa Bahasa Indonesia Spesifik Budaya. *Litera: Jurnal Penelitian Bahasa, Sastra, dan Pengajarannya*, 20(3), 415-428.
- Pamungkas, D.A., & Utami, S.S. (2022). Studi Korpus Paralel: Pemetaan Padanan Idiom Figuratif Bahasa Indonesia ke Bahasa Inggris. *Metalingua: Jurnal Penelitian Bahasa*, 20(1), 23-38.
- Pratama, F.A., & Sihombing, R.O. (2023). Rancangan Sistem Reranking Berbasis Kolokasi untuk Meningkatkan Akurasi Mesin Penerjemah Statistik. *Jurnal Teknologi Informasi*, 19(1), 1-12.
- Putra, I.G.N.S., & Dewi, P.C. (2021). Linguistik Korpus di Indonesia: Tren, Tantangan, dan Aplikasi dalam Pengembangan Teknologi Bahasa. *Kandai: Jurnal Bahasa dan Sastra*, 17(2), 299-314.
- Rahayu, W., & Hadian, T. (2022). Model Penilaian Kualitas Terjemahan Mesin dengan Metode Human Assessment Berdasarkan Aspek Akurasi dan Kewajaran. *Jurnal Edukasi dan Penelitian Informatika*, 8(3), 345-351.