

Variasi Kalimat Dialek Saudi dalam Dataset MADAR sebagai Sumber Pembelajaran Mendalam

Fatima El Zahraa^{1*}

¹Pendidikan Bahasa Arab, Fakultas Ilmu Tarbiyah dan Keguruan,
Universitas Islam Negeri Syarif Hidayatullah Jakarta, Tangerang Selatan, Banten

* Penulis korespondensi: fatimaelzahraa@uinjkt.ac.id

Abstrak

Keterbatasan sumber belajar yang memuat keragaman teks tuturan autentik dalam bentuk korpus menjadi tantangan dalam pembelajaran berbasis data. Studi ini menganalisis variasi kalimat antara dialek Jeddah dan Riyadh menggunakan dataset MADAR yang mencakup 2.000 entri untuk tiap-tiap dialek. Pendekatan kuantitatif digunakan melalui analisis statistik deskriptif dan inferensial untuk menelaah panjang kalimat dan tingkat keformalan leksikal. Hasilnya menunjukkan bahwa dialek Riyadh memiliki kalimat yang sedikit lebih panjang dan tingkat keformalan yang lebih tinggi dibandingkan dialek Jeddah. Korelasi antara panjang kalimat dan keformalan leksikal teridentifikasi lemah namun signifikan secara statistik, dengan pola lebih kentara dalam dialek Riyadh. Studi ini mengaitkan temuan tersebut dengan konsep pembelajaran mendalam yang merupakan kebalikan dari *surface-level processing* sebagai pendekatan pembelajaran, bukan dalam konteks *machine learning* dan kecerdasan buatan. Korpus tersebut dapat mendorong keterampilan analitis dan interpretatif dalam pembelajaran yang berkesadaran, bermakna, menggembirakan, dan memuliakan. Hasil analisis menyoroti koineisasi di Arab Saudi dan menyertakan perspektif bahwa tingkat keformalan yang lebih tinggi cenderung dikaitkan dengan kalimat yang lebih panjang dan ketergantungan yang lebih rendah terhadap konteks. Studi ini menunjukkan potensi integrasi variasi dialek Arab dalam pembelajaran berbasis data dan diversifikasi sumber belajar.

Kata Kunci: Dataset MADAR, Pembelajaran Mendalam, Variasi Dialek Arab

1. Pendahuluan

MADAR (*Multi-Arabic Dialect Applications and Resources*) (Habash, Nizar et al., n.d.) merupakan hasil kolaborasi antara Carnegie Mellon University Qatar, New York University Abu Dhabi, Columbia University, dan University of Bahrain. Proyek tersebut menghasilkan korpus paralel yang mencakup dialek 25 (dua puluh lima) kota—termasuk Jeddah dan Riyadh sebagai representasi dialek Arab Saudi—serta sistem otomatis untuk identifikasi dialek. Korpus tersebut disusun berdasarkan terjemahan terkendali dari korpus BTEC (*Basic Traveling Expression Corpus*), dan mencakup teks-teks yang telah disejajarkan dengan dialek Arab,

bahasa Arab standar modern (*Modern Standard Arabic/MSA*), bahasa Inggris, dan bahasa Prancis.

Korpus tersebut menggunakan CODA (*Conventional Orthography for Dialectal Arabic*), yaitu ortografi baku yang mempertahankan ciri khas dialek sambil menjaga kemiripan dengan bahasa Arab standar (*MSA*) dan CAPHI (*CAMEL Arabic Phonetic Inventory*), sistem fonetik yang merekam variasi pengucapan antardialek. Dalam alur pemrosesan bahasa alami (*Natural Language Processing/NLP*), teks dialek mengalami normalisasi dan tokenisasi, diikuti dengan ekstraksi fitur (misalnya karakter *n-gram*, *embeddings*), serta klasifikasi menggunakan model seperti SVM (*Support Vector Machine*), CNN (*Convolutional Neural Network*), atau LSTM (*Long Short-Term Memory*). Tugas anotasi—termasuk pelabelan morfologis, lemmatisasi, dan pemberian glosarium—didukung oleh antarmuka MADARi dan diperkuat melalui analisis otomatis yang dihasilkan oleh sistem MADAMIRA (Obeid et al., 2018).

Secara hukum, perlindungan terhadap set data (dataset) didasarkan pada hukum kekayaan intelektual, khususnya hak cipta dan hak basis data *sui generis* sebagaimana diatur dalam Directive 96/9/EC (The European Parliament and the Council, 1996). Sebagai hasil dari proses penerjemahan dan anotasi manual, dataset MADAR memenuhi kriteria sebagai karya orisinal dan didistribusikan dengan lisensi non-komersial, sehingga penggunaannya tunduk pada asas *pacta sunt servanda* (Aust, 2013).

Meskipun tidak memuat data pribadi, bentuk-bentuk linguistik yang melekat secara kultural dalam dataset MADAR memunculkan isu etika, seperti persoalan persetujuan, stereotip dan relasi kuasa dalam pemanfaatan ulang. Hal ini seiring dengan teori *contextual integrity* dari Nissenbaum, yang menekankan bahwa penggunaan data harus menghormati norma-norma sosial yang berlaku dalam konteks asalnya (Nissenbaum, 2004).

MADAR (Bouamor et al., 2018; Bouamor, Hassan, & Habash, 2019; Obeid et al., 2018) menunjukkan fleksibilitas penelitian dialek Arab yang telah berkembang dari pendekatan deskriptif menuju analisis *big data* dan pemrosesan mendalam. Materi autentik (Gilmore, 2007)—yang penggunaannya telah dimulai sejak karya Henry Sweet pada abad ke-19 (Sweet, 1877), namun semakin mendapat perhatian setelah berkembangnya konsep kompetensi komunikatif oleh Hymes (Hymes, 1972) sebagai pelengkap terhadap kompetensi linguistik Chomsky (Chomsky, 1965)—memegang peran penting dalam pembelajaran bahasa karena mendorong terjadinya komunikasi yang kontekstual dan bermakna.

Meskipun definisi tentang keautentikan bervariasi (Breen, 1985; Kramsch, 2012; van Lier, 2013; Widdowson, 1998), penggunaan bahasa nyata yang diarahkan kepada audiens sesungguhnya tetap menjadi inti. Sejumlah studi menunjukkan bahwa buku teks kerap terkendala merepresentasikan fitur bahasa yang autentik secara utuh (Altman, 2017). Karena itu, penting untuk memilih teks autentik yang relevan, merancang tugas-tugas berbasis konteks, dan mempertimbangkan latar budaya pembelajar (Feng & Byram, 2002). Penggunaan materi autentik mendukung pemahaman yang lebih mendalam terhadap variasi dialektal, sementara analisis kalimat memungkinkan interpretasi kontekstual dan refleksi kritis terhadap norma-norma linguistik.

Tingkat formalitas, sebagaimana didefinisikan oleh Heylighen dan Dewaele, merujuk pada sejauh mana suatu bahasa mengurangi ketergantungan pada konteks melalui struktur yang eksplisit dan terlepas dari situasi tertentu. Register formal cenderung menggunakan lebih banyak nomina, adjektiva, preposisi, dan artikel, sedangkan register informal lebih mengandalkan bentuk-bentuk yang bergantung pada konteks seperti pronomina dan verba (Dewaele & Heylighen, 2002; Heylighen & Dewaele, 1999). Dalam bahasa Arab, perbedaan tersebut sejalan dengan kontras antara bahasa Arab standar (*MSA*) dan dialek regional (Zahraa, 2024) seperti dialek Jeddah dan Riyadh. Karena itu, contoh autentik dalam korpus MADAR dapat dikaji dari segi tingkat formalitas serta implikasi sosiolinguistiknya.

Secara pedagogis, penelitian ini menggunakan dataset MADAR sebagai media untuk mendukung pembelajaran mendalam (*deep learning*)—bukan dalam konteks *machine learning* atau pun kecerdasan buatan, melainkan sebagai pendekatan pembelajaran yang mendorong keterlibatan kognitif. Pembelajaran mendalam dalam arti ini merujuk pada kemampuan siswa untuk membangun pemahaman bermakna, mengaitkan informasi baru dengan struktur pengetahuan yang sudah ada, dan menerapkan pengetahuan dalam konteks yang berbeda, bukan sekadar menghafal fakta permukaan (*surface-level processing*) (Marton & Säljö, 1976). Dataset MADAR, yang berisi variasi tuturan autentik dalam dialek Arab, memberikan peluang bagi pembelajar untuk menganalisis data nyata, membandingkan struktur kebahasaan antardialek, serta merefleksikan makna sosial-budaya dari penggunaan bahasa. Proses ini menumbuhkan keterampilan berpikir kritis seperti analisis, evaluasi, dan sintesis—ciri khas dari pembelajaran mendalam.

Dalam kerangka *big-deep-smart data*, Kalinin et al. menekankan pentingnya pemanfaatan dataset besar yang disertai pemahaman mendalam untuk menghasilkan pengetahuan yang kontekstual dan relevan (Kalinin, Sumpter, & Archibald, 2015). Hal ini selaras dengan gagasan *pervasive knowledge* oleh Anshari et al., yaitu pembelajaran lintas konteks yang berkesinambungan dan berbasis data nyata (Anshari, Alas, & Guan, 2016). Sharma et al. menekankan bahwa pemanfaatan data yang tepat dapat memperkaya pengalaman belajar daring secara adaptif (Sharma, Mahesh, & Bhuvana, 2021), sementara Dahdouh et al. menunjukkan bahwa manajemen data yang baik sangat penting untuk menciptakan pembelajaran yang bermakna dan berkelanjutan (Dahdouh, Dakkak, Oughdir, & Messaoudi, 2018). Dengan demikian, penggunaan dataset MADAR tidak hanya memperkaya konten linguistik, namun juga menciptakan konteks belajar yang mendorong pemahaman mendalam, reflektif, dan interdisipliner—baik dalam dimensi kebahasaan maupun pedagogis.

Dalam konteks pembelajaran bahasa, Godwin-Jones dan Yanhui memaparkan bahwa *big data* memungkinkan terciptanya pembelajaran yang dipersonalisasi, autentik, dan kontekstual (Godwin-Jones, 2017, 2021; Yanhui, 2018). Kessler menyoroti pentingnya pembelajaran autentik yang dimediasi oleh teknologi (Kessler, 2018), sementara Williamson memberikan peringatan terhadap potensi dominasi komersial terhadap data pendidikan (Williamson, 2017). Chen et al. serta Sedkaoui & Khelfaoui berpendapat bahwa tantangan utama dalam penerapan *big data* dalam pendidikan terletak pada pemahaman mekanisme belajar untuk mengoptimalkan nilai edukatif dari data tersebut (Chen, Yin, Isaias, & Psotka, 2020; Sedkaoui & Khelfaoui, 2018). Karena itu, Buyya et al. mendorong pemanfaatan *big data* untuk belajar dari masa lalu, memahami situasi saat ini, dan mempersiapkan diri menghadapi masa depan (Buyya, Calheiros, & Dastjerdi, 2016).

Selaras dengan perkembangan tren *EdTech* (Zahraa, Fatima El et al., 2024), dalam pembelajaran bahasa Arab tercatat adanya peningkatan signifikan dalam penelitian terkait pembelajaran berbantuan perangkat *mobile*, dengan penekanan pada pentingnya adaptasi terhadap keragaman budaya-linguistik serta perhatian terhadap isu etika dan aksesibilitas (Alasmari, 2025), sebagaimana tercatat pula adanya keterkaitan antara pemanfaatan teknologi dengan kontekstualisasi dalam proses pembelajaran bahasa Arab (Keshav, Julien, & Miezal, 2022; Qamarya, Nurul et al., 2023; Sarwani et al., 2025; Zahraa, 2025). Sementara Al-Ayyoub et al. menekankan urgensi penerapan *deep learning* dalam pemrosesan bahasa alami untuk

bahasa Arab, mengingat keterbatasan sumber daya yang tersedia (Al-Ayyoub, Nuseir, Alsmearat, Jararweh, & Gupta, 2018).

Eksposur terhadap ekspresi autentik sangat berarti untuk memahami variasi bahasa. Namun, apabila data tersebut diperlakukan sebagai contoh semata, pembelajar mungkin akan melewatkan kedalaman makna kultural yang dikandungnya. Studi ini merespons permasalahan tersebut dengan memajukan variasi kalimat kontekstual dalam dialek Saudi, yang dibatasi pada dialek Jeddah dan Riyadh yang hanya terdapat dalam dataset MADAR, dengan tujuan mendorong kesadaran terhadap keberagaman tanpa mengandaikan adanya imersi yang mendalam.

Tinjauan pustaka mengenai dialek Jeddah dan Riyadh mengungkap sejumlah perbedaan mendasar. Dialek Riyadh, yang merupakan bagian dari kelompok Najdi sedentari, memperoleh status semi-formal secara nasional karena kedudukannya yang sentral secara politik. Al-Rojaie menunjukkan bahwa dialek Najdi, khususnya varian Riyadh, memainkan peran penting dalam proses koineisasi nasional—yaitu pembentukan dialek standar campuran yang dalam konteks Arab Saudi menjadikan dialek Riyadh sebagai pusatnya (Al-Rojaie, 2023). Dialek para pekerja non-Saudi juga memiliki relasi saling memengaruhi dengan kedua dialek Saudi tersebut (Zahraa, Nawawi, & Muttaqin, 2024).

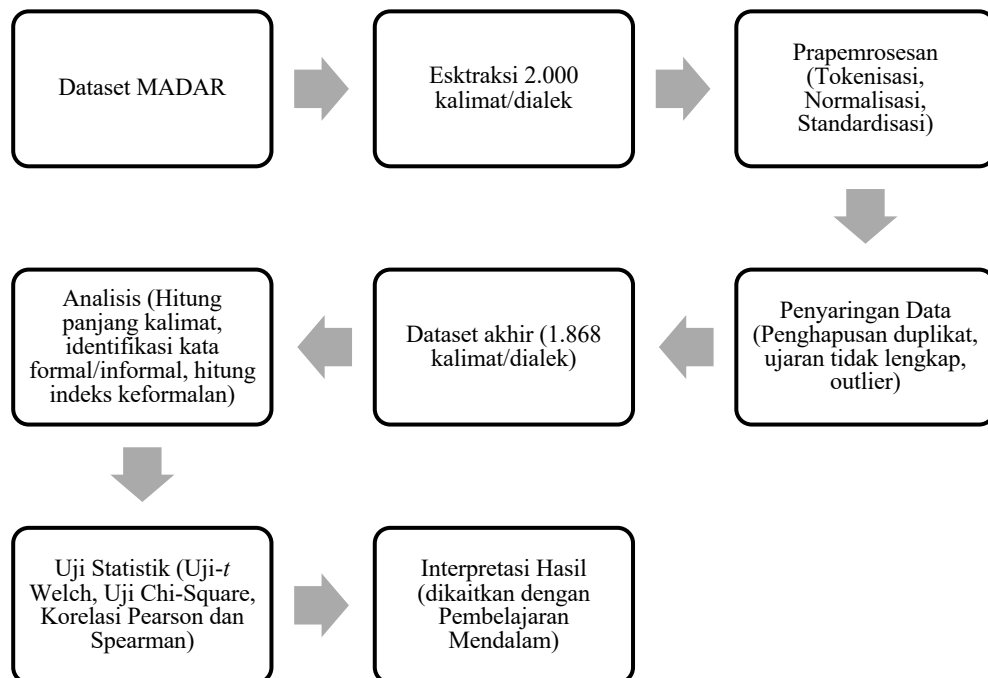
Ikrish et al. menemukan bahwa dialek Riyadh memiliki metrik ritmis khas (%V, ΔV), yang menunjukkan struktur kalimat lebih panjang dan kompleks (Ikrish, Alotaibi, & Sandouka, 2020). Sementara itu, dialek Hijazi urban yang digunakan di Jeddah mencerminkan akar kosmopolitan (Omar, 1975). Literatur mencatat ciri khas dialek tersebut lebih informal dan ringkas, sebagai contoh, kata kerja terduplikasi seperti *yfalfil* digunakan untuk memadatkan makna (Kabli, 2023). Ikrish et al. juga mencatat bahwa dialek Mekkah—yang memiliki kedekatan dengan dialek Jeddah—menunjukkan tempo bicara lebih cepat dan persentase vokal (%V) yang lebih tinggi, mencerminkan gaya tutur yang lebih singkat dan langsung (Ikrish et al., 2020).

Sejumlah studi telah mengkaji kedua dialek tersebut, namun demikian masih terbuka ruang untuk mengkomparasikan keduanya dalam lingkup dialek Saudi sebagai bagian dari variasi bahasa Arab. Studi ini berupaya memberikan tinjauan awal sederhana terhadap variasi kalimat dengan mengeksplorasi data dari korpus MADAR.

2. Metode

Studi ini menggunakan dataset MADAR dengan fokus pada variasi dialektal Arab Saudi. Dataset seimbang yang terdiri dari 2.000 kalimat awal diekstraksi masing-masing untuk dialek Jeddah dan Riyadh. Entri-entri tersebut merepresentasikan interaksi tutur alami, seperti pertanyaan santai atau permintaan bantuan, misalnya *يَمْدِيكَ تَدُور مَعَايَةَ عَلَى شَنْطَتِي؟* (Jeddah) dan *تَقْدِر تَسَاعِدُنِي بِالْعَثُور عَلَى بَدَلَتِي؟* (Riyadh). Proses penyaringan data secara sistematis menghasilkan dataset akhir sebanyak 1.868 kalimat representatif untuk tiap-tiap dialek.

Prosedur analisis dalam penelitian ini dilakukan melalui sejumlah tahap seperti tampak pada Gambar 1. Tahap pertama, prapemrosesan data dengan tokenisasi, normalisasi, dan standardisasi untuk memastikan konsistensi bentuk bahasa. Tahap kedua melibatkan analisis leksikal, yaitu setiap kata dikategorikan sebagai formal atau informal dengan mempertimbangkan perspektif Heylighen dan Dewaele. Tahap ketiga, analisis statistik terhadap data yang telah diproses, menggunakan uji-*t* Welch, uji Chi-Square, serta korelasi Pearson dan Spearman. Tahap keempat, interpretasi hasil analisis mengacu pada kerangka pembelajaran mendalam (Kemdikdasmen, 2025).



Gambar 1. Diagram Alur Analisis Data

3. Hasil dan Pembahasan

3.1 Panjang Kalimat dalam Dialek Jeddah dan Riyadh

Sebanyak 3.736 kalimat dianalisis dari dataset MADAR yang telah disortir, yang terbagi secara seimbang antara dialek Jeddah ($n = 1.868$) dan dialek Riyadh ($n = 1.868$). Rata-rata panjang kalimat pada dialek Riyadh adalah 5,71 kata ($SD = 1,34$), sementara dialek Jeddah menunjukkan rata-rata yang lebih pendek, yaitu 5,44 kata ($SD = 1,29$). Uji- t Welch menunjukkan perbedaan yang signifikan secara statistik antara kedua dialek, $t(3728,61) = 6,27$, $p < 0,0000000004$.

Ukuran efek yang dihitung dengan Cohen's $d = 0,205$ menunjukkan kecenderungan kecil namun konsisten bahwa tuturan dialek Riyadh mengandung kalimat yang lebih panjang. Meskipun perbedaan ini tergolong kecil dalam hal besar pengaruh, hasil ini mengisyaratkan adanya kemungkinan perbedaan dalam struktur wacana atau preferensi sintaksis antara dua wilayah dialek tersebut.

3.2 Tingkat Keformalan Antardialek

Proporsi item leksikal formal menunjukkan perbedaan yang cukup kentara antara kedua dialek. Dalam dialek Riyadh, sebesar 4,60% dari keseluruhan kata diklasifikasikan sebagai formal, sedangkan dalam dialek Jeddah proporsinya jauh lebih rendah, yaitu 2,17%. Uji Chi-Square independensi dilakukan terhadap distribusi item leksikal formal dan non-formal di antara kedua dialek berdasarkan data agregat dari 3.736 kalimat.

Hasil analisis menunjukkan signifikansi statistik, $\chi^2(1, N = 3.736) = 8,509$, $p = 0,0035$, yang mengonfirmasi adanya hubungan yang tidak bersifat acak antara jenis dialek dan tingkat keformalan leksikal. Nilai *odds ratio* yang diperoleh ($OR = 2,19$) menunjukkan bahwa kata-kata formal boleh jadi lebih dari dua kali muncul dalam tuturan dialek Riyadh dibandingkan dengan dialek Jeddah. Nilai ekspektasi di tiap kategori sudah memenuhi syarat minimum, yang berarti hasil uji statistik valid secara teknis.

Hasil ini menguatkan interpretasi bahwa tuturan dialek Riyadh cenderung memuat leksikon formal, yang kemungkinan mencerminkan norma-norma sosiolinguistik terkait identitas regional, kesadaran terhadap register (Zahraa, 2024; Zahraa et al., 2024), atau pengaruh institusional. Hasil kuantitatif ini sejalan dengan kesan kualitatif dari studi-studi dialektologi sebelumnya, yang kerap mengaitkan tuturan dialek Riyadh dengan kedekatan yang lebih tinggi terhadap *MSA* dalam konteks publik atau semi-formal. Sebaliknya, rendahnya proporsi kosakata formal dalam tuturan dialek Jeddah dapat mencerminkan gaya interaksi yang lebih kolokial dan informal, sebagaimana lazim ditemukan dalam komunitas tutur perkotaan

pesisir. Kontras ini juga konsisten dengan koineisasi (Al-Rojaie, 2023), yaitu kontak antardialek dan perataan atau pelandaian linguistik di kota seperti Riyadh dan Jeddah turut membentuk pola baru dalam tingkat keformalan dan penggunaan bahasa.

3.3 Korelasi Antara Panjang Kalimat dan Tingkat Keformalan

Untuk mengkaji hubungan antara panjang kalimat dan tingkat keformalan, dilakukan analisis korelasi Pearson dan Spearman menggunakan gabungan dataset sebanyak 3.736 kalimat. Koefisien korelasi Pearson menunjukkan nilai $r = 0,061$, $t(3734) = 3,73$, $p = 0,00019$, yang mengindikasikan adanya hubungan linear yang signifikan secara statistik, meskipun lemah, antara kedua variabel tersebut.

Sementara itu, koefisien korelasi peringkat Spearman menunjukkan nilai yang lebih tinggi, yaitu $\rho = 0,227$, $t(3734) = 14,24$, $p < 7,27 \times 10^{-45}$, yang menunjukkan adanya hubungan monotonik yang cukup signifikan. Koefisien Spearman yang lebih kuat mengisyaratkan adanya pola non-linear, yaitu kalimat yang lebih panjang cenderung memuat lebih banyak item leksikal formal, terutama dalam dialek Riyadh.

Hal ini juga diperkuat dengan pengamatan sebaran data pada diagram pencar (Gambar 2), yang menunjukkan adanya pengelompokan nilai keformalan yang lebih tinggi pada kalimat-kalimat panjang dalam dataset dialek Riyadh. Sebaliknya, nilai pada dialek Jeddah tampak lebih padat dan cenderung ke tingkat keformalan yang lebih rendah. Hasil ini menyiratkan bahwa panjang kalimat bukanlah satu-satunya prediktor kuat terhadap tingkat keformalan leksikal. Faktor kontekstual dan pragmatik kemungkinan besar turut memengaruhi variasi keformalan antardialek.

3.4 Skor Keformalan Berdasarkan Indeks Heylighen dan Dewaele

Skor yang dihitung menggunakan indeks Heylighen dan Dewaele menunjukkan tingkat keformalan linguistik yang tinggi pada kedua dialek. Dialek Jeddah memperoleh skor sebesar 99,96, sementara dialek Riyadh sedikit lebih rendah, yaitu 99,92. Uji- t Welch terhadap kedua skor tersebut menunjukkan perbedaan yang signifikan secara statistik, $t = 7,82$, $p < 0,0001$. Namun demikian, selisih absolut sebesar 0,04 poin tergolong sangat kecil dan dinilai memiliki implikasi praktis yang terbatas.

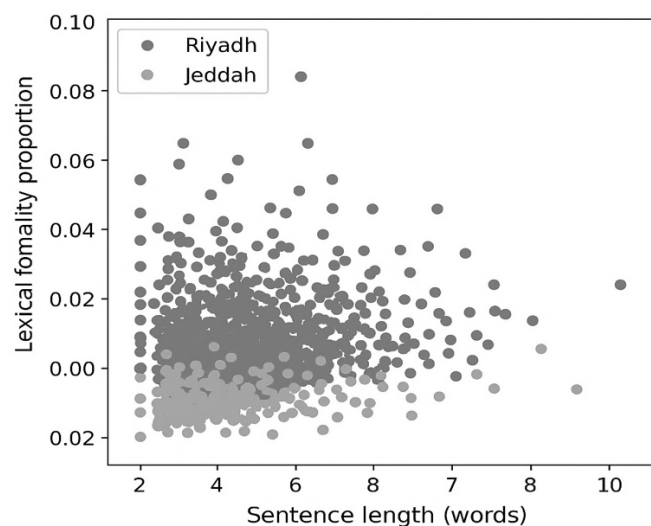
Secara keseluruhan, hasil studi ini menunjukkan bahwa keformalan tidak bersifat seragam antarindikator. Meskipun dialek Riyadh menunjukkan kecenderungan lebih formal berdasarkan panjang kalimat dan proporsi leksikon formal, skor indeks Heylighen dan Dewaele

justro sedikit lebih tinggi pada dialek Jeddah. Perbedaan ini mengindikasikan bahwa keformalan bersifat multidimensional dan mungkin dipengaruhi oleh faktor-faktor lain seperti genre wacana, konteks penggunaan, dan ciri sintaksis non-leksikal. Karena itu, pendekatan analisis kualitatif dan kontekstual tetap diperlukan untuk memahami ekspresi keformalan secara lebih utuh dan mendalam.

3.5 Ringkasan Hasil Statistik

Tabel 1. Perbandingan Hasil Statistik antara Dialek Jeddah dan Riyadh

Parameter	Jeddah	Riyadh	Uji Statistik	p-value
Rata-rata panjang kalimat	5,44	5,71	$t(3728,61) = 6,27$	$p < 0,0000000004$
Proporsi kata formal (%)	2,17%	4,60%	$\chi^2(1) = 8,509$	$p = 0,0035$
Korelasi panjang kalimat– keformalan (Pearson)	$r = 0,061$	—	$t(3734) = 3,73$	$p = 0,00019$
Korelasi panjang kalimat– keformalan (Spearman)	$\rho = 0,227$	—	$t(3734) = 14,24$	$p < 7,27 \times 10^{-45}$
Skor keformalan (rata-rata, Heylighen & Dewaele)	99,96	99,92	$t(3734) = 7,82$	$p < 0,0001$



Gambar 2. Diagram Pencar Panjang Kalimat dan Tingkat Keformalan

Sebagaimana ditampilkan pada Gambar 2, titik-titik data dari dialek Riyadh menunjukkan penyebaran yang lebih luas serta kecenderungan naik secara moderat, yang mengindikasikan bahwa kalimat yang lebih panjang cenderung mengandung proporsi kata formal yang lebih

tinggi. Sebaliknya, dialek Jeddah menunjukkan distribusi yang lebih terkonsentrasi, dengan panjang kalimat yang lebih pendek dan tingkat keformalan leksikal yang lebih rendah.

Pola visual tersebut mendukung hasil statistik dari analisis korelasi Pearson dan Spearman, yang meskipun menunjukkan hubungan yang lemah antara panjang kalimat dan tingkat keformalan, memperlihatkan bahwa asosiasi tersebut lebih tampak pada dialek Riyadh. Dengan demikian, diagram pencar tersebut memberikan konfirmasi visual terhadap perbedaan dialektal, serta memperkuat interpretasi bahwa keformalan linguistik tidak tersebar merata di antara varietas regional.

3.6 Implikasi Pedagogis

Bagian ini menempatkan hasil penelitian dalam kerangka pendidikan konstruktivis (Vygotsky & M. Cole, 2018) dan responsif budaya. Perbedaan linguistik—terlihat pada panjang kalimat dan tingkat keformalan—tidak hanya bersifat deskriptif, namun juga membuka ruang bagi kesadaran kritis terhadap bahasa (Fairclough, 1992), kompetensi antarbudaya, dan pembelajaran inklusif. Dalam pandangan sosiolinguistik, dialek merupakan penanda identitas dan afiliasi sosial (Holmes & Wilson, 2022), yang dapat mendorong pebelajar untuk merefleksikan variasi bahasa. Perbandingan fitur dialektal memungkinkan kritik terhadap asumsi seperti bahwa kompleksitas struktur selalu berarti keformalan. Proses ini mendukung pembelajaran berbasis bukti dan kemandirian melalui tugas kontekstual yang bermakna guna meningkatkan kompetensi komunikatif (Canale & Swain, 1980).

Pemanfaatan dataset dialektal untuk pembelajaran didukung oleh penelitian sebelumnya. Younes menunjukkan bahwa integrasi dialek lokal dalam pembelajaran bahasa Arab sebagai bahasa asing dapat mendorong keterlibatan dan pemahaman terhadap variasi sosial bahasa (Younes, 2014). Sementara itu, Abdulrahman membuktikan bahwa tugas berbasis korpus efektif dalam meningkatkan kompetensi pragmatik dan refleksi kebahasaan (Abdulrahman, 2018).

Penerapan nilai-nilai berkesadaran, bermakna, menggembirakan, dan memuliakan dalam pembelajaran mendalam berbasis dataset dialektal dapat diwujudkan sebagai berikut. Pertama, refleksi berkesadaran (Sullivan, 2020) mendorong pebelajar untuk menelaah respons pribadi dan emosional (Dray & Wisneski, 2011) terhadap bentuk-bentuk dialektal. Kedua, tugas bermakna (Frey & Fisher, 2010; McHugh, 2016)—seperti menyusun dialog beranotasi dari interaksi autentik—memperkuat keterlibatan linguistik dan kontekstualisasi budaya. Ketiga,

aktivitas menggembarakan, seperti gamifikasi (misalnya tantangan dialek/تحدي اللهجات), berpotensi meningkatkan motivasi dan retensi kognitif (Costello, 2020; Smirani & Yamani, 2024) melalui keterlibatan fisik (olah raga) (Merino-Campos, 2025), di samping melibatkan olah pikir, olah rasa, olah hati. Keempat, proyek bercerita yang memuliakan keberagaman (Cooper, 1994; Craig, Hull, Haggart, & Crowder, 2001; T. Scott Bledsoe & Kimberly A. Setterlund, 2020) dapat memvalidasi ekspresi dialektal, serta menumbuhkan rasa kepemilikan dan apresiasi terhadap diversitas kebudayaan.

Praktik pedagogis tersebut dilandaskan pada visi pendidikan holistik yang sejalan dengan delapan dimensi profil lulusan, yaitu (1) keimanan dan ketakwaan (Q.S. Al-Hujurat:13; Ar-Rum:22), (2) kewargaan (Republik Indonesia, 1945; UNESCO, 2015), (3) penalaran kritis (Bloom, 1956), (4) kreativitas (seperti melalui tugas-tugas divergen) (Nisred Gasharov, Heyrullah Mahmudov, Dibirasulav Nurmagomedov, Elmira Ramazanova, & Nasrudin Magomedov, 2019), (5) kolaborasi (misalnya melalui Google Colab) (Carneiro et al., 2018), (6) kemandirian (Hu & Yeo, 2020), (7) kesehatan (seperti *mental health* (Han, Zhang, Tao, & Zhao, 2021)), dan (8) komunikasi dalam konteks antarbudaya (Wei, 2024).

Selain itu, integrasi wawasan dialektal membuka peluang untuk pengembangan aplikasi berbasis *Natural Language Processing (NLP)*, seperti klasifikasi kalimat peka dialek (Sayadi, A. A., Abulohoom, A., Uddin, S. A., Lataifeh, M., & Elnagar, 2021), penambahan diakritik otomatis (harakat/*tashkīl*) (Masmoudi, Mdhaftar, Sellami, & Hadrich, 2019; Zayyan, Elmahdy, Husni, & Al, 2016), pelabelan kelas kata (*part-of-speech tagging*) (AbuZeina, Wasfi Al-Khatib, & Al-Muhtaseb, 2011), pengukuran kompleksitas sintaksis, pemodelan konteks semantik, dan sebagainya. Berbagai alat dapat mendukung fungsi pedagogis dan analitis, terutama dalam pembelajaran bahasa Arab digital yang menjembatani register formal dan kolokial.

Penelitian ini memiliki beberapa keterbatasan. Proses prapemrosesan masih bersifat sederhana dan deterministik, sehingga kurang fleksibel dalam menangani variasi bahasa alami. Penilaian keformalan sangat bergantung pada kualitas dan cakupan daftar kata, sedangkan pengukuran panjang kalimat dipengaruhi oleh teknik segmentasi yang digunakan. Studi ini juga belum melibatkan model *NLP* modern seperti BERT atau AraBERT yang terbukti unggul dalam berbagai tugas kebahasaan. Aspek prosodik dan pragmatik belum dibahas, dan data hanya bersumber dari korpus MADAR, tanpa melibatkan korpus lain seperti NADI, IADD, DART, atau ADI17 yang dapat memberikan representasi lebih luas.

Untuk penelitian selanjutnya, disarankan memperluas ukuran sampel, menggunakan analisis yang lebih kompleks (seperti regresi atau multivariat), serta menambahkan validasi, triangulasi, dan pengujian reliabilitas. Pendekatan kualitatif, seperti analisis wacana, juga perlu dijajaki untuk menangkap aspek pragmatik dan ideologis yang tidak terjangkau secara statistik. Selain itu, pemanfaatan *machine learning* dengan fitur linguistik dapat meningkatkan akurasi klasifikasi tingkat keformalan dan memperluas kontribusi dalam kajian linguistik komputasional atau lainnya.

4. Simpulan

Studi ini menunjukkan bahwa variasi dialek Saudi dalam dataset MADAR memberikan wawasan terhadap dimensi linguistik dan sosiokultural. Kalimat dalam dialek Riyadh cenderung lebih panjang dan tampak lebih formal secara struktural. Namun demikian, dialek Jeddah menunjukkan tingkat keformalan yang sedikit lebih tinggi menurut indeks Heylighen & Dewaele. Korelasi antara panjang kalimat dan tingkat keformalan, meskipun ada, tidak cukup kuat untuk dijadikan indikator tunggal —menunjukkan bahwa keformalan merupakan fenomena kompleks yang tidak hanya ditentukan oleh panjang kalimat. Korelasi yang relatif rendah mengindikasikan bahwa keformalan linguistik bersifat kontekstual dan tidak dapat direduksi menjadi metrik struktural semata. Hasil studi menyoroti integrasi data dialektal ke dalam pembelajaran bahasa Arab guna mendorong pembelajaran mendalam. Selaras dengan nilai-nilai pendidikan abad ke-21, dataset MADAR tidak hanya berperan sebagai korpus linguistik, namun juga sebagai sarana untuk menumbuhkan kesadaran kritis, kompetensi antarbudaya, dan komunikasi yang inklusif.

Daftar Pustaka

- Abdulrahman, R. A. I. M. (2018). *Role of Corpus-Based Pragmatic Explicit Instructions in Developing Inter language Competence* (Al-Neelain University). Al-Neelain University. Retrieved from <https://repository.neelain.edu.sd/items/cda8eff7-cdb2-4549-b27c-d1ae73fa5318>
- AbuZeina, D., Wasfi Al-Khatib, M. E., & Al-Muhtaseb, H. (2011). Toward Enhanced Arabic Speech Recognition Using Part of Speech Tagging. *International Journal of Speech Technology*, 14, 419–426. <https://doi.org/https://doi.org/10.1007/s10772-011-9121-5>

- Al-Ayyoub, M., Nuseir, A., Alsmearat, K., Jararweh, Y., & Gupta, B. (2018). Deep Learning for Arabic NLP: A Survey. *Journal of Computational Science*, 26, 522–531. <https://doi.org/10.1016/j.jocs.2017.11.011>
- Al-Rojaie, Y. (2023). Sociolinguistics in Saudi Arabia: Present Situation and Future Directions. *Journal of Arabic Sociolinguistics*, 1(1), 76–97. <https://doi.org/10.3366/arabic.2023.0006>
- Alasmari, T. (2025). Artificial Intelligence and M-Learning in Arabic Countries: Innovations, Trends, and Regional Perspectives. *International Journal of Interactive Mobile Technologies*, 19(5), 170–194.
- Altman, S. (2017). Reinterpreting the Right to An Open Future: From Autonomy to Authenticity. *Law and Philosophy*, 37(4), 415–436. <https://doi.org/10.1007/s10982-017-9317-1>
- Anshari, M., Alas, Y., & Guan, L. S. (2016). Developing Online Learning Resources: Big Data, Social Networks, and Cloud Computing to Support Pervasive Knowledge. *Education and Information Technologies*, 21, 1663–1677. <https://doi.org/10.1007/s10639-015-9407-3>
- Aust, A. (2013). Vienna Convention on the Law of Treaties 1969. In *Modern Treaty Law and Practice*. <https://doi.org/10.1017/cbo9781139152341.005>
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives The Classification of Educational Goals*. Longmans.
- Bouamor, H., Habash, N., Salameh, M., Zaghoulani, W., Rambow, O., Abdulrahim, D., ... Oflazer, K. (2018). The MADAR Arabic Dialect Corpus and Lexicon. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 3387–3396. Retrieved from <http://www.ustar-consortium.com/>
- Bouamor, H., Hassan, S., & Habash, N. (2019). The MADAR Shared Task on Arabic Fine-Grained Dialect Identification. *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 199–207. Retrieved from <http://resources.camel-lab.com>.
- Breen, M. P. (1985). Authenticity in the Language Classroom. *Applied Linguistics*, 6(1), 60–70. <https://doi.org/10.1093/applin/6.1.60>
- Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (2016). Big Data: Principle & Paradigms. In *Morgan Kaufmann*.
- Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, 1(1), 1–47.

<https://doi.org/10.1093/applin/I.1.1>

- Carneiro, T., Nóbrega, R. V. M. Da, Nepomuceno, T., Bian, G.-B., Albuquerque, V. H. C. De, & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, 6, 61677–61685. <https://doi.org/10.1109/ACCESS.2018.2874767>
- Chen, N. S., Yin, C., Isaias, P., & Psotka, J. (2020). Educational Big Data: Extracting Meaning from Data for Smart Education. *Interactive Learning Environments*, 28(2), 142–147. <https://doi.org/10.1080/10494820.2019.1635395>
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.
- Cooper, C. S. (1994). *Storytelling in the Basic Course for the Promotion of Cultural Diversity*. New Orleans.
- Costello, R. (2020). *Gamification Strategies for Retention, Motivation, and Engagement in Higher Education: Emerging Research and Opportunities*. IGI Global.
- Craig, S., Hull, K., Haggart, A. G., & Crowder, E. (2001). Storytelling Addressing the Literacy Needs of Diverse Learners. *TEACHING Exceptional Children*, 33(5), 46–51. <https://doi.org/10.1177/004005990103300507>
- Dahdouh, K., Dakkak, A., Oughdir, L., & Messaoudi, F. (2018). Big Data for Online Learning Systems. *Education and Information Technologies*, 23(6), 2783–2800. <https://doi.org/10.1007/s10639-018-9741-3>
- Dewaele, J. M., & Heylighen, F. (2002). Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7, 293–340.
- Dray, B. J., & Wisneski, D. B. (2011). Mindful Reflection as a Process for Developing Culturally Responsive Practices. *TEACHING Exceptional Children*, 44(1), 28–36. <https://doi.org/10.1177/004005991104400104>
- Fairclough, N. (1992). Discourse and Text: Linguistic and Intertextual Analysis within Discourse Analysis. *Discourse & Society*, 3(2), 193–217. <https://doi.org/10.1177/0957926592003002004>
- Feng, A., & Byram, M. (2002). Authenticity in College English Textbooks - an Intercultural Perspective. *RELC Journal*, 33(2), 58–84.
- Frey, N., & Fisher, D. (2010). Motivation Requires a Meaningful Task. *English Journal*, 100(1), 30–36. <https://doi.org/10.2307/20787688>

- Gilmore, A. (2007). Authentic Materials and Authenticity in Foreign Language Learning. *Language Teaching*, 40(2), 97–118. <https://doi.org/10.1017/S0261444807004144>
- Godwin-Jones, R. (2017). Emerging Technologies: Scaling Up and Zooming In: Big Data and Personalization in Language Learning. *Language Learning and Technology*, 21(1), 4–15.
- Godwin-Jones, R. (2021). Big Data and Language Learning: Opportunities and Challenges. *Language Learning and Technology*, 25(1), 4–19.
- Habash, Nizar, Kemal Oflazer, Houda Bouamor, Owen Rambow, Mohammed Salameh, Wajdi Zaghoulani, ... Sabit Hassan. (n.d.). The MADAR Project. Retrieved May 19, 2025, from <https://sites.google.com/nyu.edu/madar/>
- Han, J., Zhang, Z., Tao, J., & Zhao, Z. (2021). Deep Learning for Mobile Mental Health: Challenges and Recent Advances. *IEEE Signal Processing Magazine*. <https://doi.org/10.1109/MSP.2021.3099293>
- Heylighen, F., & Dewaele, J. (1999). Formality of Language: Definition, Measurement and Behavioral Determinants. In *Internet Bericht* (Vol. 1999).
- Holmes, J., & Wilson, N. (2022). *An Introduction to Sociolinguistics, Sixth Edition*. Routledge. <https://doi.org/10.4324/9780367821852>
- Hu, X., & Yeo, G. B. (2020). Emotional Exhaustion and Reduced Self-Efficacy: The Mediating Role of Deep and Surface Learning Strategies. *Motivation and Emotion*, 785–795. <https://doi.org/10.1007/s11031-020-09846-2>
- Hymes, D. (1972). On Communicative Competence In: J.B. Pride and J. Holmes. (Eds) *Sociolinguistics. Selected Readings*, pp. 269–293.
- Ikriş, R. F. B., Alotaibi, Y. A., & Sandouka, S. B. (2020). Extended Speech Rhythm-Based Analysis of Saudi Dialects Using SAAVB Corpus. *Proceedings of the 2020 14th International Conference on Innovations in Information Technology, IIT 2020*, 172–177. <https://doi.org/10.1109/IIT50501.2020.9299065>
- Kabli, H. M. (2023). Reduplication in Urban Hijazi Arabic Dialect Spoken in the Western Region of Saudi Arabia. *English Linguistics Research*, 12(1), 1–11. <https://doi.org/10.5430/elr.v12n1p1>
- Kalinin, S. V., Sumpter, B. G., & Archibald, R. K. (2015). Big-Deep-Smart Data in Imaging for Guiding Materials Design. *Nature Materials*, 14, 973–980. <https://doi.org/10.1038/nmat4395>

- Kemdikdasmen. (2025). *Pembelajaran Mendalam*. Pusat Kurikulum dan Pembelajaran Badan Standar, Kurikulum, dan Asesmen Pendidikan Kementerian Pendidikan Dasar dan Menengah Republik Indonesia.
- Keshav, M., Julien, L., & Miezal, J. (2022). The Role Of Technology In Era 5.0 In The Development Of Arabic Language In The World Of Education. *Journal International of Lingua and Technology*, 1(2), 79–98. <https://doi.org/10.55849/jiltech.v1i2.85>
- Kessler, G. (2018). Technology and the Future of Language Teaching. *Foreign Language Annals*, 51(1), 205–218. <https://doi.org/10.1111/flan.12318>
- Kramersch, C. (2012). Authenticity and Legitimacy in Multilingual SLA. *Critical Multilingualism Studies*, 1(1), 107–128.
- Marton, F., & Säljö, R. (1976). On Qualitative Differences in Learning: I—Outcome and Process. *British Journal of Educational Psychology*, 46(1), 4–11. <https://doi.org/10.1111/j.2044-8279.1976.tb02980.x>
- Masmoudi, A., Mdhaftar, S., Sellami, R., & Hadrich, L. (2019). Automatic Diacritics Restoration for Tunisian Dialect. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3), 1–18. <https://doi.org/https://doi.org/10.1145/3297278>
- McHugh, M. C. (2016). Experiencing Flow: Creativity and Meaningful Task Engagement for Senior Women. *Women and Therapy*, 39(3–4), 280–295. <https://doi.org/10.1080/02703149.2016.1116862>
- Merino-Campos, C. (2025). Enhancing Physical Education Through Gamification and Ergonomics: A Literature Review. *Theoretical and Applied Ergonomics*, 1(1), 1–12. <https://doi.org/https://doi.org/10.3390/tae1010003>
- Nisred Gasharov, Heyrullah Mahmudov, Dibirasulav Nurmagomedov, Elmira Ramazanova, & Nasrudin Magomedov. (2019). Divergent Tasks as a Means of Developing the Creativity of Younger Students. *Atlantic Press*, 267–271. <https://doi.org/10.2991/mplg-ia-19.2019.50>
- Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, 79(1), 119–157. Retrieved from <https://nyuscholars.nyu.edu/en/publications/privacy-as-contextual-integrity>
- Obeid, O., Khalifa, S., Habash, N., Bouamor, H., Zaghouani, W., & Oflazer, K. (2018).

- MADARi: A Web Interface for Joint Arabic Morphological Annotation and Spelling Correction. *ArXiv Preprint ArXiv:1808.08392*, 2616–2622.
- Omar, M. K. (1975). *Saudi Arabic Urban Hijazi Dialect*. Foreign Service Institute.
- Qamarya, Nurul, Andi Fitriani Djollong, Siti Mukaromah, Ely Widayati, Adi Isma, Rabiyyatul Adawiyah, ... Edy Siswanto. (2023). *Model Pembelajaran*. Eureka Media Aksara.
- Republik Indonesia. (1945). *Undang Undang Dasar Negara Republik Indonesia Tahun 1945 (yang dipadukan dengan perubahan I, II, III dan IV)*. Retrieved from https://bphn.go.id/data/documents/uud_1945.pdf
- Sarwani, S., Maria Susan, Ariya Pannaditthana Candra, Fatimah Setiani, Fatima El Zahraa, Veni Nella Syahputri, ... Sabrina. (2025). *Bahasa dan Literasi dalam Pendidikan Modern Pendekatan Inovatif dan Keterampilan Abad 21*. Eureka Media Aksara.
- Sayadi, A. A., Abulohoom, A., Uddin, S. A., Lataifeh, M., & Elnagar, A. (2021). Region-Level Arabic Dialect Identification Using Deep Learning Models. *IET Conference Proceedings*, 288–301. The Institution of Engineering and Technology; The 2nd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2021). <https://doi.org/https://doi.org/10.1049/icp.2021.2683>
- Sedkaoui, S., & Khelfaoui, M. (2018). Understand, Develop and Enhance the Learning Process with Big Data. *Information Discovery and Delivery*, 47(1), 2–16. <https://doi.org/10.1108/IDD-09-2018-0043>
- Sharma, K., Mahesh, T. R., & Bhuvana, J. (2021). Big Data Technology for Developing Learning Resources. *Journal of Physics: Conference Series*, 1979(1). <https://doi.org/10.1088/1742-6596/1979/1/012019>
- Smirani, L., & Yamani, H. (2024). Analysing the Impact of Gamification Techniques on Enhancing Learner Engagement, Motivation, and Knowledge Retention: A Structural Equation Modelling Approach. *Electronic Journal of E-Learning*, 22(9), 111–124. <https://doi.org/https://doi.org/10.34190/ejel.22.9.3563>
- Sullivan, T. N. (2020). *Mindful Reflection: Does Intentional Reflection Enhance Learner Creativity and Innovation?* (Northcentral University). Northcentral University. Retrieved from <https://www.proquest.com/openview/707d5caa3cb8f981a9454c74ea14c90a/1?pq-origsite=gscholar&cbl=51922&diss=y>
- Sweet, H. (1877). *Handbook of Phonetics*. Clarendon Press.

- T. Scott Bledsoe, & Kimberly A. Setterlund. (2020). *Using Narratives and Storytelling to Promote Cultural Diversity on College Campuses Advances in Higher Education and Professional Development*. IGI Global.
- The European Parliament and the Council. (1996). Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases. *Official Journal of the European Communities*, 20–28. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009>
- UNESCO. (2015). Global Citizenship Education: Topics and Learning Objectives. In *Global citizenship education: topics and learning objectives*. <https://doi.org/10.54675/drhc3544>
- van Lier, L. (2013). *Interaction in the Language Curriculum: Awareness, Autonomy, and Authenticity*. Routledge. <https://doi.org/10.4324/9781315843223>
- Vygotsky, L., & M. Cole. (2018). Lev Vygotsky: Learning and Social Constructivism. *Learning Theories for Early Years Practice*, 68–73.
- Wei, T. (2024). The Role of English in Promoting International Cultural Exchange in the Context of Big Data and Deep Learning. *Journal of Computational Methods in Sciences and Engineering*, 24, 369–384. <https://doi.org/10.3233/JCM-237021>
- Widdowson, H. G. (1998). Context, Community, and Authentic Language. *TESOL Quarterly*, 32(4), 705–716.
- Williamson, B. (2017). Who Owns Educational Theory? Big Data, Algorithms and the Expert Power of Education Data Science. *E-Learning and Digital Media*, 14(3), 1–18. <https://doi.org/10.1177/2042753017731238>
- Yanhui, W. (2018). Language E-learning Based on Learning Analytics in Big Data Era. *Proceedings of the 2018 International Conference on Big Data and Education*, 106–111. <https://doi.org/10.1145/3206157.3206177>
- Younes, M. (2014). *The Integrated Approach to Arabic Instruction*. Routledge. <https://doi.org/https://doi.org/10.4324/9781315740614>
- Zahraa, Fatima El, Muhamad Kosim Gifari, Rabiyyatul Adawiyah, Iyam Maryati, Nurma Yunita Peu'uma, Nandang Kusmana, ... Muhammad Arsyam. (2024). *Perencanaan Pendidikan: Konsep dan Langkah Strategis*. Eureka Media Aksara.
- Zahraa, F. El. (2024). Sociolinguistic Competence and Sociolinguistic Appropriateness in the Context of Arabic Language Based on the Common European Framework of Reference

- for Languages Scale. *Al-Ittijah: Jurnal Keilmuan Dan Kependidikan Bahasa Arab*, 16(2), 1–24. <https://doi.org/10.32678/alittijah.v16i2.10882>
- Zahraa, F. El. (2025). Leveraging Artificial Intelligence and Digital Technologies to Enhance Sociolinguistic Competence and Arabic Language Skills. *International Collaborative Conference on Multidisciplinary Science*, 8–19. International Forum of Researchers and Lecturers. <https://doi.org/https://doi.org/10.70062/iccms.v2i1.65>
- Zahraa, F. El, Nawawi, M., & Muttaqin, Z. (2024). Sociolinguistic Content in Training Courses: Pidgin Arabic and White Dialect for Bus Drivers. *Tanwir Arabiyyah: Arabic As Foreign Language Journal*, 4(2), 211–232. <https://doi.org/https://doi.org/10.31869/aflj.v4i2.6012>
- Zayyan, A. A., Elmahdy, M., Husni, H., & Al, J. M. (2016). Automatic Diacritics Restoration for Dialectal Arabic Text. *International Journal of Computing & Information Science*, 12(2), 159–165. <https://doi.org/http://dx.doi.org/10.21700/ijcis.2016.119>