At-Ta'dib. Vol. 17. No. 1, June 2022 Available online at:
ISSN: 0216-9142 https://ejournal.unida.gontor.ac.id/index.php/tadib/index
DOI: http://dx.doi.org/10.21111/at-tadib.v17i1.7067 e-ISSN: 2503-3514

Korpus Arab Pesantren: Digitizing the work of Arabic non-Arabic speakers at Modern Islamic Institution Darussalam Gontor

Yoke Suryadarma
Universitas Darussalam Gontor
yoke.suryadarma@unida.gontor.ac.id

Gamal Abdul Nasir Zakaria
Universiti Brunei Darussalam
gamal.zakaria@ubd.edu.bn

Received December 14, 2021/Accepted May 23, 2022

Abstract

Modern Islamic Institution Darussalam Gontor (PMDG) is One of the educational institutions in Indonesia that has been consistent in learning Arabic since its inception due to there are many Arabic works produced by the student community component within the institution. The purpose of this study is to specifically describe the works of non-Arabic Speaker Arabic students and explain the process of collecting and digitizing Arabic corpus data sources made by non-Arabic Speaker Arabic students at Modern Islamic Institution Darussalam Gontor, Ponorogo, East Java. This research employs a field study with a descriptive qualitative analysis approach as a research method. The results reveal that, first, the works of Arabic language students based on the type of data can be categorized into two parts, namely written Arabic works (written products) and spoken Arabic works (Spoken Products). Conventional written data can be read by software and conventional data cannot be read. Second, the process of collecting and digitizing Arabic corpus data by Arabic non-Arabic Speaker students at Modern Islamic Institution Darussalam Gontor was carried out in three stages. The first stage converts the conventional data in the form of handwriting and voice recordings to digital data in **doc* format. The second stage is the conversion of digital data in **doc* format into plain text *(*txt)* format. The third stage is to enter data in **txt* format into a web-based data processing engine, namely Sketch Engine. Thus, the digital corpus data is ready to be processed into a study based on certain linguistic objectives.

Keywords : Korpus Arab, Digitizing the work of Arabic, Gontor, data korpus, *Sketch Engine*

Korpus Pesantren Arab: Digitalisasi Karya Penutur Arab Non-Arab di Pondok Modern Darussalam Gontor

Pendahuluan

Praktik pembelajaran dan penggunaan bahasa Arab telah lama dilakukan oleh masyarakat Indonesia terutama kalangan pelajar muslim di berbagai institusi pendidikan Islam mulai tingkat dasar, menengah dan atas baik formal maupun nonformal. Salah satu lembaga yang konsisten dalam pembelajaran bahasa Arab sejak awal berdirinya adalah Pondok Modern Darussalam Gontor (PMDG).

Terdapat banyak karya-karya bahasa Arab yang dihasilkan oleh komponen masyarakat pelajar di dalam lembaga tersebut, baik itu individu santri maupun komunitas santri, yang dilakukan di dalam dan di luar kelas, baik karya berupa tulisan maupun ucapan. Hal ini karena PMDG telah berhasil menciptakan lingkungan berbahasa Arab yang kondusif dan mileu belajar yang baik dalam pembelajaran bahasa Arab.[1] Karena itulah, PMDG dapat dijadikan sebagai sumber material korpus bahasa Arab yang unik, karena bahasa Arab tersebut diucapkan dan ditulis oleh orang Indonesia asli dan berada jauh dari kawasan Arab.

Korpus sendiri merupakan kajian linguistik yang secara khusus meneliti bahasa melalui seperangkat data yang bersifat alamiah, riil sesuai penggunaannya, baik itu data tulisan maupun data lisan yang ditranskripsikan.[2] Menurut pengertiannya, korpus merupakan sekumpulan data, baik data biasa maupun data digital, dalam bentuk tertulis yang berisi bermacam-macam informasi kebahasaan, mulai dari tataran kata, struktur, makna, dan wacana, yang dapat dimanfaatkan untuk penelitian.[3] Bertolak dari sini, sangat memungkinkan dilakukan suatu identifikasi karya bahasa Arab dan proses digitalisasinya suatu sumbangan penting dan berharga untuk memulai langkah strategis dalam penyusunan korpus bahasa Arab di wilayah Asia Tenggara, khususnya Indonesia bahkan bisa melangkah lebih luas menjadi kawasan Nusantara yang melebar dari Malaysia, Thailand, Brunei Darussalam dan Philipina.

Penelitian mengenai kajian Korpus bahasa Arab di di Indonesia belum begitu banyak. Adapun beberapa kajian mengenai korpus Arab di Indonesia diantaranya. Penelitian dengan judul, "Linguistik korpus dalam kajian dan pembelajaran bahasa Arab di Indonesia" yang ditulis oleh Nur Hizbullah, dkk (2016) memberikan informasi mengenai linguistik korpus serta perkembangan dan dinamikanya dalam kajian bahasa Arab di dunia internasional, mengenai kemungkinan penyusunan suatu model korpus bahasa Arab di Indonesia sekaligus pemanfaatannya dalam kajian dan pembelajaran

---

[1] Abdul Hafidz Zaid, "تكنولوجيا التعليم المقترحة لتعليم مهارة الكلام لطلاب المستوى المتوسط في إندونيسيا," *LISANUDHAD* 1, no. 2 (December 8, 2014), accessed November 7, 2020, https://ejournal.unida.gontor.ac.id/index.php/lisanu/article/view/446.

[2] Svenja Adolphs, *Introducing Electronic Text Analysis A Practical Guide for Language and Literary Studies*, 1st ed. (New York: Routledge, 2006), 137.

[3] Nur Hizbullah, Fazlur Rachman, and Fuzi Fauziah, "Linguistik Korpus Dalam Kajian Dan Pembelajaran Bahasa Arab Di Indonesia," in *Konferensi Nasional Bahasa Arab (KONASBARA) II*, 2016, 385.

bahasa Arab di tingkat perguruan tinggi.[4] Di akhir penelitian peneliti memberikan saran kepada para ahli bahasa Arab di Indonesia ikut menyumbangkan kepada dunia kajian bahasa Arab global suatu model korpus bahasa Arab dengan mengangkat khazanah sumber pustaka pembelajaran bahasa Arab dan sumber lain produk ilmiah asli putra-putri negeri ini.

Penelitian selanjutnya dengan judul, *"Projected Characteristics and Content of Arabic Corpus in Indonesia"*, yang ditulis oleh Nur Hizbullah & Muchlis, M. (2018) memberikan hasil mengenai klasifikasi model korpus-korpus yang telah berkembang di dunia dan pemanfatannya dalam rangka mewujudkan korpus Arab di Indonesia. Adapun metode yang dipakai adalah analis dekriptif.[5] Ada juga penelitian dengan judul, *"Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia"* yang ditulis oleh Nur Hizbullah, dkk (2019) menyimpulkan bahwa bahan material yang berpotensi dijadikan korpus pembelajar bahasa Arab di pesantren bisa didapatkan melalui tiga jenis kegiatan, yaitu kegiatan formal-kurikuler, nonformal/ekstrakurikuler, dan informal serta subkegiatan yang tercakup di dalamnya. Korpus pembelajar yang terdapat di pesantren sangatlah luas, beragam, dan spesifik. Namun demikian, baru sedikit produk pembelajaran yang sudah berbentuk data digital, sementara sebagian besarnya masih berupa data konvensional, baik berupa lisan maupun tulisan.[6] Melihat luas dan beragamnya data kebahasaan Arab di pesantren, terbuka peluang dan tantangan untuk mengeksplorasi lebih jauh situasi dan kondisi riil pembelajaran bahasa Arab melalui penelitian multidisipliner berbasis korpus.

Dari ketiga penelitian terdahulu diatas, terlihat bahwa kajian terdahulu pertama dan kedua membahas mengenai keberadaan korpus Arab sebagai kajian yang layak untuk dikaji dan pemanfaatannya dalam kajian pembelajaran bahasa Arab di Indonesia. Sehingga keduaya tidak memiliki kesamaan yang berarti dengan penelitian yang akan diteliti oleh peneliti. Sedangkan kajian terdahulu ketiga membahas tentang identifikasi sumber-sumber materi dan data yang ada di beberapa pesantren modern termasuk Pondok Modern darussalam Gontor (PMDG) hanya saja kajian yang dilakukan di dalamnya masih terlalu umum dan belum spesifik, karena peneliti hanya membagi hasil penemuan dari identifikasi sumber-sumber materi dan data korpus Arab kepada tiga elemen saja, selain itu peneliti dalam penelitian tersebut juga menyimpulan bahwa proses digitalisasi dari hasil temuan tersebut belum dilakukan sehingga menyarankan agar dilakukan penelitian lebih lanjut dalam hal ini.

[4] Hizbullah, Rachman, and Fauziah, "Linguistik Korpus Dalam Kajian Dan Pembelajaran Bahasa Arab Di Indonesia."

[5] Nur Hizbullah and Muchlis Madian Muhammad, "Projected Characteristics and Content of Arabic Corpus in Indonesia," *Advances in Social Science, Education and Humanities Research (ASSEHR)* 154, no. Icclas 2017 (2018): 172–174.

[6] Nur Hizbullah et al., "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia," *KnE Social Sciences*, no. July (2019): 3–4.

Penelitian selanjutnya dilakukan oleh Azzahra, dkk (2020) mengenai pemanfaatan korpus linguistik dalam pengembangan kamus kedokteran.[7] Hasil penelitian menunjukan bahwa dengan bantuan Anchonc, software pengolah data korpus berhasil menyusun kamus kedokteran yang terdiri dari 400 kosakata bersumber dari sepuluh website berita bahasa Arab. Penelitian terkait dengan kegunaan data korpus digital untuk menyusun kamus juga dilakukan oleh Suryadarma & Alinda (2020), dimana hasil penelitian tersebut adalah tersusunya kamus Az-Ziro'ah sebagai media ajar pembelajaran Bahasa Arab yang terdiri dari 3374 kosakata yang berhubungan dengan pertanian.[8] Dari kedua penelitian ini, terlihat bahwa data korpus digital dapat dimanfaatkan untuk berbagi kepentingan dan tujuan tertentu, seperti pembuatan kamus dwibahasa.

Dari sini dapat terlihat bahwa penelitian yang akan dilakukan oleh peneliti sekarang berbeda secara substantif dengan kelima penelitian tersebut, dan dapat dikatakan sebagai penelitian lanjutan yang bertumpu pada hasil dari penelitian tersebut dimana beberapa sumber data korpus Arab telah terindentifikasi, namun belum terjadi proses digitalisasi data. Dengan demikian, penelitian yang peneliti lakukan ini akan fokus pada penyempurnaan pengidentifikasian sumber data korpus Arab berupa karya-karya bahasa Arab yang dihasilkan oleh para pelajar bahasa Arab *non arabic speaker* Pondok Modern Darussalam Gontor (PMDG) baik itu sumber korpus bahasa Arab tertulis dan terucap. dan menambahnya dengan melakukan observasi lapangan secara langsung, korespondesi dengan pihak terkait dan telaah dokumentasi dari sumber-sumber data korpus Arab yang ada di lapangan.

Kedua, melakukan upaya penghimpunan sumber data korpus Arab dalam satu wadah, dan ketiga adalah melakukan upaya digitalisasi sumber-sumber data korpus Arab tersebut ke dalam komputer dengan memakai aplikasi Ms. Word sampai akhirnya menghasilkan karya digital berupa korpus pelajar sehingga dapat diolah lebih lanjut untuk keperluan basis data dalam suatu aplikasi pengolah data korporasi, seperti korpus website dan konkordansi Arab. Bertolak dari hal ini, maka jelas tujuan penelitian ini adalah mendeskripsikan secara spesifik karya-karya pelajar bahasa Arab *non Arabic Speaker* dan menjelaskan proses penghimpunan dan proses digitalisasi sumber data korpus Arab karya-karya pelajar bahasa Arab *non Arabic Speaker* di Pondok Modern Darussalam Gontor Ponorogo Jawa Timur.

Metode Penelitian

Jenis penelitian ini merupakan penelitian lapangan (*field study*) yang menitikberatkan pada tempat dimana peneliti melakukan penelitian. Adapun metode penelitian menggunakan pendekatan deskriptif analisis kualitatif. Menurut sugiono,

[7] Siti Fatimah Azzahra, Nur Hizbullah, and Iin Suryaningsih, "Penyusunan Kamus Kedokteran Arab – Indonesia Dengan Pendekatan Linguistik Korpus," *Tsaqofiya : Jurnal Pendidikan Bahasa dan Sastra Arab* 2, no. 2 (2020): 60–66.

[8] Yoke Suryadarma and Alinda Zakiyatul Fakhiroh, "Tashmīm Qāmus 'al-Zirā'Ah' Kawasīlah Ta'allum Al-'Arabiyyah Li Thalabah Qism Al-Tiknūlūjiyā Al-Shinā'iyyah Al-Zirā'iyyah Muassasn 'Alā Al-Mudawwanah Al-Lughowiyyah," *LISANUDHAD* 7, no. 2 (December 17, 2020): 37–56, accessed October 20, 2021, https://ejournal.unida.gontor.ac.id/index.php/lisanu/article/view/6744.

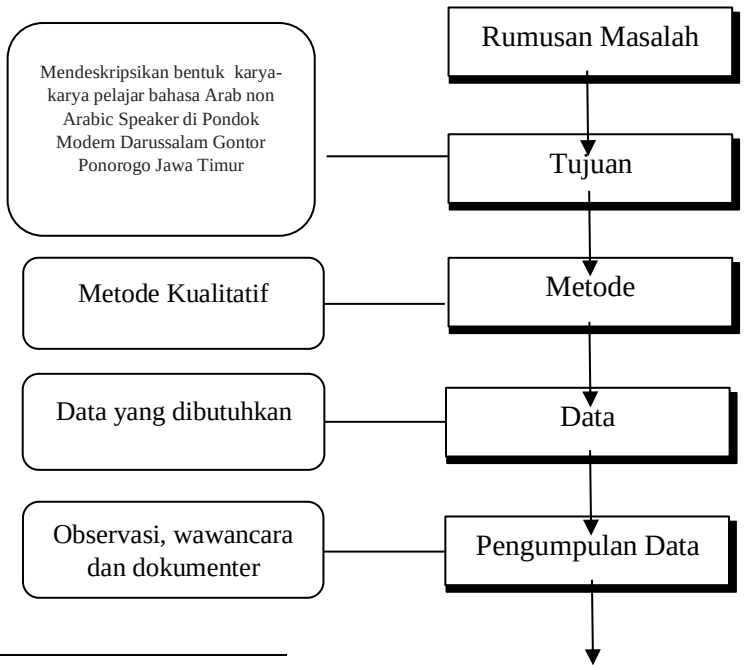
metode penelitian ini berlandaskan pada filsafat pospositivisme, digunakan untuk meneliti pada kondisi obyek yang alamiah, dimana peneliti adalah sebagai instrumen kunci, teknik dilakukan secara trianggulasi (gabungan beberapa teknik), analisis data bersifat induktif alias kualitatif, dan hasil penelitian lebih menekankan makna dari pada generalisasi. [9]

Data-data dalam metode kualitatif ini dikumpulkan dengan menggunakan metode trianggulasi (*trianggulation*) yang menggabungkan metode observasi, wawancara dan documenter [10]. Metode observasi, peneliti lakukan dengan terjun langsung ke lapangan guna mengetahui dan melihat secara lebih dalam berbagai kegiatan yang dilakukan oleh para pelajar bahasa Arab *non arabic speaker* dan produk-produk atau karya-karya yang dihasilkan oleh mereka baik berbentuk tulisan maupun lisan.

Kemudian metode wawancara/interview yaitu dengan cara mendatangi dan bertanya secara langsung kepada beberapa orang yang terlibat dan mempunyai pengaruh dalam penelitian ini.[11] Dan terakhir metode dokumenter yaitu, peneliti menelurusi berbagai dokumentasi guna mendapatkan tambahan informasi yang berhubungan dengan penelitian yang sedang diteliti oleh peneliti.[12]

Adapun untuk menganalisis data, peneliti menggunakan analisis data model Miles and Huberman, melalui tiga tahapan analisis data; display data, data reduksi dan penarikan kesimpulan.[13] Berikut bagan kerangka pemikiran penelitian :

Bagan 1: Kerangka Pemikiran Penelitian



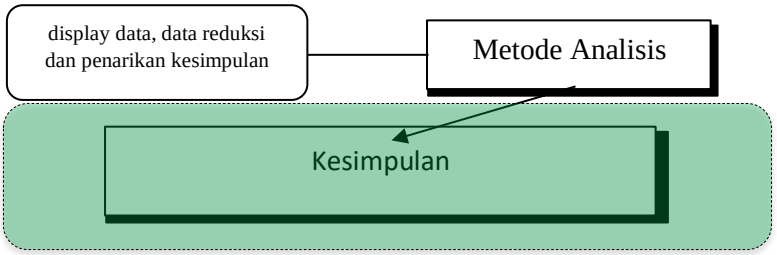
[9] Sugiyono, *Metode Penelitian Dan Pengembangan (Research and Development)*, 4th ed. (Bandung: Alfabeta, 2019), 9.

[10] Sugiyono, *Metode Penelitian Pendidikan*, Cetakan 27. (Bandung: CV. Alfabeta, 2018), 129.

[11] Emzir, *Metodologi Penelitian Kualitatif : Analisis Data*, 6th ed. (Depok: Rajawali Pres, 2018), 50.

[12] Ibid., 61.

[13] Ibid., 129.



Pembahasan

Korpus dalam pembelajaran bahasa Arab merupakan kajian baru yang terus ditumbuhkembangkan. Korpus sendiri terdiri dari banyak ragam. Nesselhauf (2011) menyebutkan bahwa ragam-ragam korpus antara lain:[14]

1. Korpus general/referensi dan—"kebalikannya"—korpus khusus.
2. Korpus historis dan—kebalikannya—korpus bahasa modern.
3. Korpus regional,
4. Korpus pembelajar dan—kebalikannya—korpus penutur asli.
5. Korpus multilingual dan—kebalikannya—korpus ekabahasa.
6. Korpus lisan dan—kebalikannya—korpus tulisan dan/atau campuran lisan-tulisan.
7. Korpus ortografis dan—kebalikannya—korpus beranotasi.

Dalam penelitian ini, peneliti memfokuskan penelitiannya pada jenis korpus pembelajar dari kalangan penutur asing (pelajar bahasa Arab dari orang asli Indonesia). Maka berbagai produk yang dihasilkan melalui proses pembelajaran bahasa Arab dapat dikategorikan sebagai korpus pembelajar. Salah satu korpus pembelajar bahasa Arab yang sudah ada sebagai contoh adalah *The Arabic Learner Corpus*.[15] Korpus ini menjaring data dari sejumlah pembelajar penutur Arab asli dan sebagian penutur asing. Korpus lainnya dalam konteks pembelajaran adalah korpus bahasa Arab anak (*Arabic Children's Corpus)* yang merupakan kompilasi dari sejumlah teks yang termuat dalam buku pelajaran maupun kisah-kisah untuk pembelajar usia dini.[16] Alfaifi dan Atwell memproyeksikan, keberadaan korpus pembelajar semacam ini bisa dimanfaatkan untuk

[14] Nadja Nesselhauf, *Corpus Linguistics : A Practical Introduction, Anglistisches Seminar* (Heidelberg: Uniheidelberg, 2011), http://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus-Linguistics-Practical -Introduction.pdf.

[15] Abdullah Alfaifi, "The Arabic Learner Corpus Website," last modified 2015, https://www.arabiclearnercorpus.com/.

[16] Latifa Al-sulaiti et al., "Compilation of an Arabic Children's Corpus," in *LREC 2016: 10th Language Resources and Evaluation Conference*, ed. Nicoletta Calzolari (Portorož, Slovenia, 2016), https://eprints.whiterose.ac.uk/100839/.

berbagai bidang riset, antara lain analisis kontrastif intrabahasa, pembuatan kamus pembelajar, pemerolehan bahasa kedua, desain materi pembelajaran, dan teknik optical character recognition (OCR).[17]

Dalam penelitian yang dilakukan oleh peneliti, terdapat tiga tahapan pembahasan, pertama, penelitian ini fokus pada penyempurnaan identifikasi dan kategorisasi sumber data korpus Arab berupa karya-karya bahasa Arab yang dihasilkan oleh para pelajar bahasa Arab *non arabic speaker* Pondok Modern Darussalam Gontor (PMDG) baik itu sumber korpus bahasa Arab tertulis dan terucap. dan menambahnya dengan melakukan observasi lapangan secara langsung, korespondesi dengan pihak terkait dan telaah dokumentasi dari sumber-sumber data korpus Arab yang ada di lapangan. Kedua, melakukan upaya digitalisasi sumber-sumber data korpus Arab tersebut ke dalam komputer. Ketiga, menyiapkan data digital ke dalam format khusus yang dapat dioleh melalui mesin pengolah korpus.

Terkait dengan tahapan pembahasan pertama, yaitu pengidentifikasian dan kategorisasi sumber data korpus Arab berupa karya-karya bahasa Arab yang dihasilkan oleh para pelajar bahasa Arab *non arabic speaker* Pondok Modern Darussalam Gontor (PMDG) baik itu sumber korpus bahasa Arab tertulis dan terucap, Peneliti telah mendapatkan setidaknya sampai detik ini 150 dokument data bersumber dari sumber dokument tertulis dan dokument yang ditranskip dari rekaman suara.

Kesemua data tersebut didapatkan dari proses Wawancara dengan beberapa walikelas, guru KMI, staf KMI, dan staf pengasuhan santri di Pondok Modern Darussalam Gontor, observasi secara langsung di beberapa kegiatan dan aktivitas santri, seperti ketika *muhadatsah shobah*, *muhadoroh*, *muhadatsah* harian, pengumuman di masjid, dan ketika proses pengajaran. Dan terakhir, pengumpulan data tersebut juga didapatkan dari hasil penelusuran dokument, seperti tulisan harian santri (*Insya' Yaumi*), tulisan tugas materi *Insya* di kelas, teks-teks pidato berbahasa Arab yang dipakai ketika latihan pidato, teks khutbah jumat, teks *i'dad amaliyah tadris*, teks *khutbatul wada'*, dan lain sebaginya. Lengkapnya dapat dilihat dari tabel berikut :

Tabel 1. Dokument berdasarkan kategori Sumber data Tertulis

No	Tertulis
1	Tulisan harian santri (*Insya' Yaumi*)
2	Tulisan tugas materi Insya di kelas
3	Teks-teks pidato berbahasa Arab
4	Teks khutbah jumat berbahasa Arab
5	Teks i'dad amaliyah tadris
6	Teks khutbatul wada'
7	Teks Paper/Tugas akhir karya ilmiah santri kelas Enam

[17] Abdullah Alfaifi and Eric Atwell, "Potential Uses of the Arabic Learner Corpus" (2013), accessed December 14, 2021, http://www.uclouvain.be/en-cecl-longdale.html.

8	Soal Ujian Materi-Materi KMI (materi bahasa Arab, materi Dirosah Islamiyah, materi Tarbiyah wa ta'lim)
9	Buku Syarhu Mahfudzhot KMI
10	Buku Muhadatasah Al-yaumiyah
11	Buku Ushul Tarbiyah
12	Buku Tarjamah kelas 1 – 4 KMI
13	Buku Mahfudzot kelas 5 KMI
14	Wardun (Majalah Tahunan PM Gontor) tahun 2021
15	Wardun (Majalah Tahunan PM Gontor) tahun 2020
16	Wardun (Majalah Tahunan PM Gontor) tahun 2019
17	Wardun (Majalah Tahunan PM Gontor) tahun 2018
18	Jawaban Materi Insya (menulis bahasa Arab) Santri
19	Teks MC bahasa Arab
20	Teks MC Pembacaan Ketetapan Hasil Musyawarah Kerja OPPM dan Koordinator
21	Teks MC LPJ dan Serah Terima Amanat
22	Teks MC Tamu Grand Syaikh Al-Azhar
23	Teks MC pembukaan ajaran baru
24	Teks Acara Khutbatul Iftitah
25	Teks Upacara
26	Berbagai jenis surat (surat undangan, surat kesediaan, surat pembuat soal, surat pengawas, surat tugas, surat izin, surat keterangan, surat kumpul koordinasi, surat pengarahan, surat rapat, dll)
27	Buku Iza'ah (pengumuman dalam berbagai jenis)
28	Naskah Drama Bahasa Arab kelas lima
29	Naskah Drama Bahasa Arab kelas enam
30	Kumpulan Naskah Drama Bahasa Arab per-Rayon dalam acara Arabic Drama Contes
31	Majalah dinding
32	Pesan nasihat sebelum liburan

Tabel 2. Dokument berdasarkan kategori Sumber data Terucap

No	**Terucap**
1	Rekaman *Muhadatsah Shobah*
2	Rekaman *muhadoroh*
3	*Muhadatsah* harian
4	Pengumuman-pengumuman di masjid / melalui pengeras suara
5	Proses pengajaran
6	*Taujihat* Asatidz

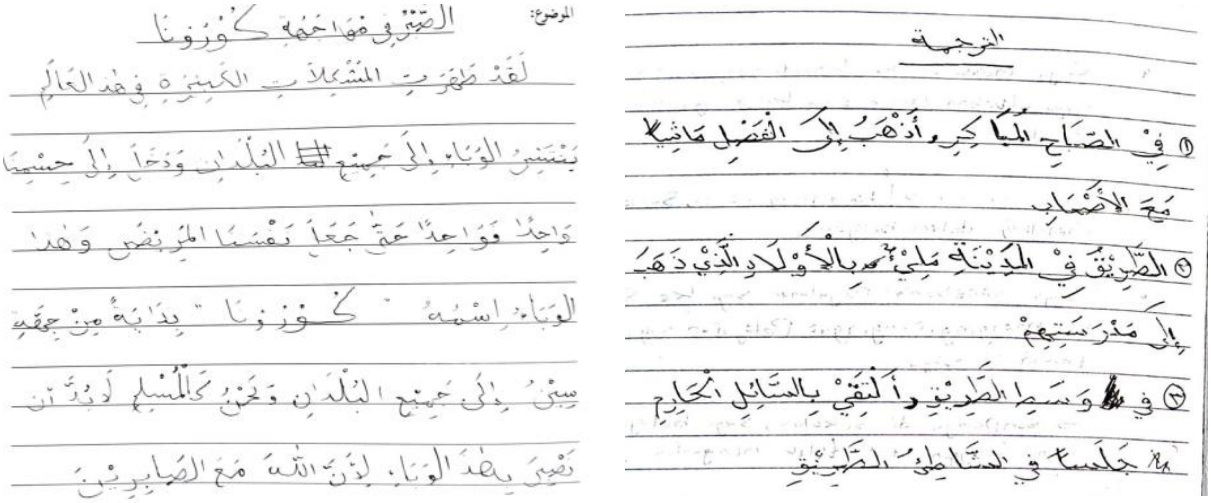
7	Rekaman Khutbah Jumat oleh kelas Enam KMI
8	Rekaman Drama Arena kelas Lima KMI
9	Rekaman Drama Kelas Enam dalam Panggung Gembira

Dari tahapan pertama diatas, maka peneliti melanjutkan pembahasanya ke tahap kedua yaitu melakukan proses digitalisasi data. Yang dimaksud dengan digitalisasi data melakukan proses konversi data konvensional ke data digital.[18] Data konvensional merupakan data asli yang masih berupa tulisan tangan ataupun berupa rekaman suara.

Untuk data konvensional berupa tulisan tangan masih perlu dikonversi ke data digital melalui proses pemindaian mengunakan suatu aplikasi atau software tertentu dan/ataupun penulisan ulang. Bagus tidaknya hasil pemindaian data konvensinal bergantung kepada keadaan riil data tersebut. Semakin data tersebut bagus, maka proses pemindaiannya akan lebih mudah begitu pula dengan proses penyuntingannya.[19] Namun jika data tersebut tidak bagus, sulit terbaca dan sebagainya, maka peneliti akan melakukan proses penulisan ulang untuk mengubah data tersebut menjadi data digital.

Berikut contoh data konvensional yang sulit terbaca dengan software dan proses digitalisasinya dilakukan menggunakan penulisan ulang :

Gamba 1,2 : Contoh data konvensional



Sedangkan data konvensional yang berupa rekaman suara, proses konversi ke data digital dilakukan secara manual, dimana peneliti menulis semua rekaman tersebut langsung ke aplikasi pengolah data, dalam hal ini peneliti menggunakan Ms. Word.

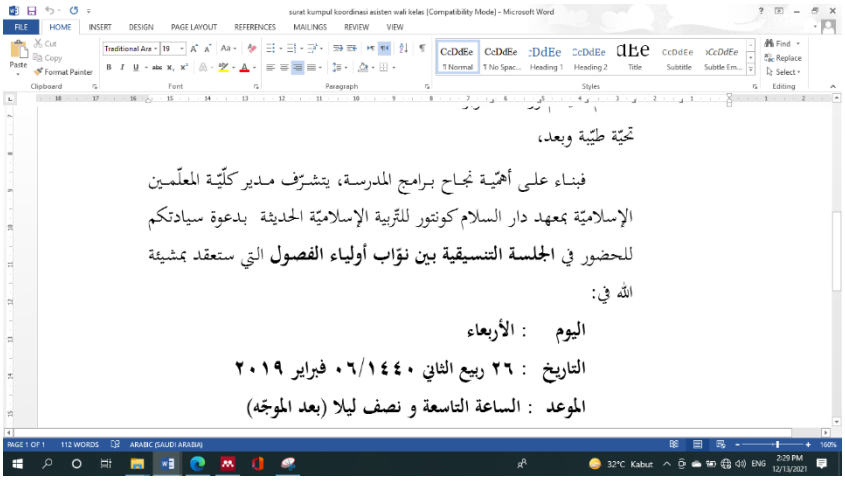
[18] Hizbullah et al., "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia."

[19] Nur Hizbullah et al., "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia," *KnE Social Sciences* 2019 (2019): 980–989, https://www.kne-publishing.com/index.php/KnE-Social/article/view/4922.

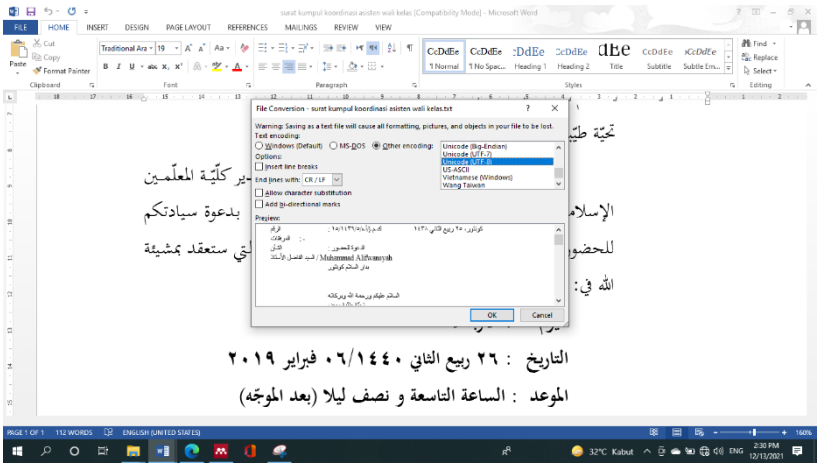
Serangkaian proses ini merupakan proses yang membutuhkan banyak waktu, tenaga dan pikiran serta meminta tingkat kesabaran yang cukup tinggi.

Adapun data material yang sudah berbentuk digital (sudah dalam aplikasi format *doc atau *rtf, dan sebagainya) termasuk data hasil konversi dari konvensional ke data digital, maka hanya memerlukan konversi ke format khusus (*txt) dan disunting lalu dibersihkan sebelum dijadikan data korpus. Berikut contoh file yang sudah berbentuk *doc yang kemudian dikonversi ke format (*txt) :

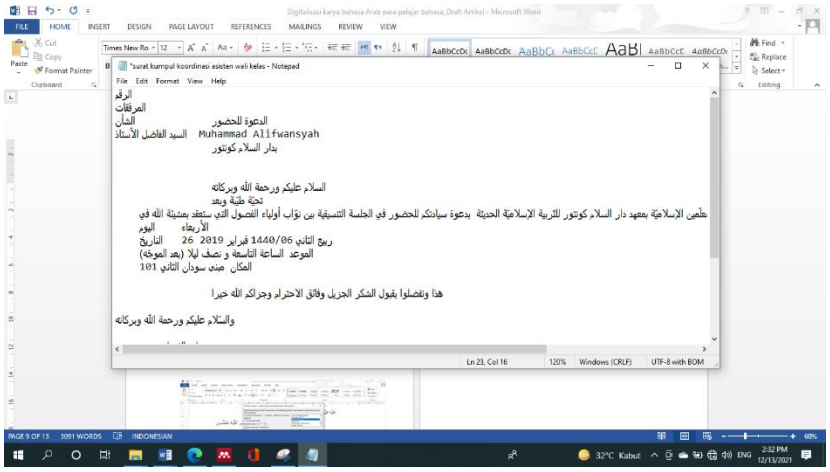
Gambar 3 : Contoh file yang sudah berbentuk *doc



Gambar 4 : proses konversi ke ke format (*txt)

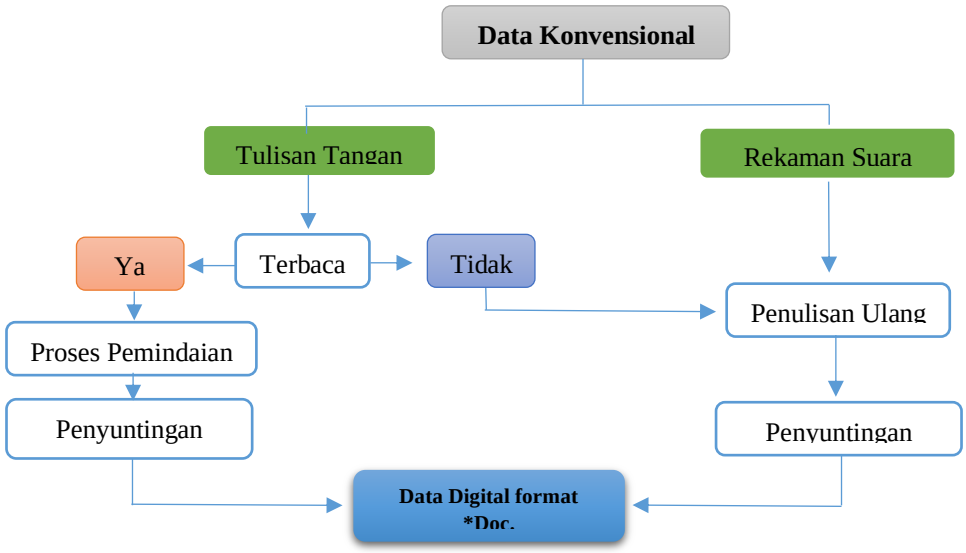


Gambar 5 : hasil proses konversi ke ke format (*txt)



Dari sini, maka proses konversi data konvensional ke data digital dilakukan sesuai dengan kondisi riil data konvensional. Berikut alur prosesnya :

Bagan 1. Alur proses konversi data konvensional ke data digital

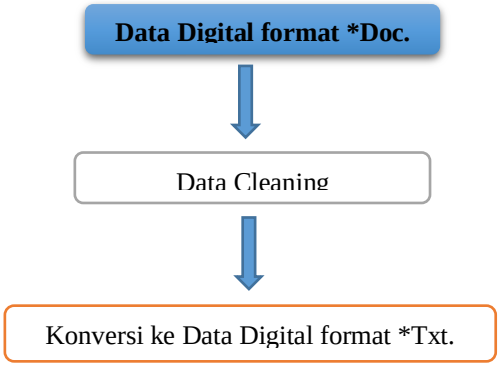


Ketika semua data sudah menjadi data digital (sudah dalam aplikasi format *doc atau *rtf), maka tahapan ketiga adalah mengubah data atau file tersebut menjadi data dengan format Plain Text (*txt). Menurut Suryadarma & Alinda (2020), Sebelum melakukan proses konversi, sebaiknya peneliti terlebih dahulu melakukan proses pengeditan dan penyuntingan dengan membuang semua unsur-unsur tulisan redaksi (tanda baca, angka, dll) yang tidak diperlukan melalui proses *data cleaning*.[20] Hal ini dilakukan agar mendapatkan hasil yang optimal di dalam proses analisis data korpus

[20] Yoke Suryadarma and Alinda Zakiyatul Fakhiroh, "Optimalisasi Penggunaan Corpus Linguistics Dalam Penyusunan Kamus Az- Ziro'ah Sebagai Media Pembelajaran Bahasa Arab," in *International Seminar on Language, Education, and Culture (ISoLEC) 2020* (Malang: Universitas Negeri Malang, 2020), 123–128, http://isolec.um.ac.id/proceeding/index.php/issn/article/view/59.

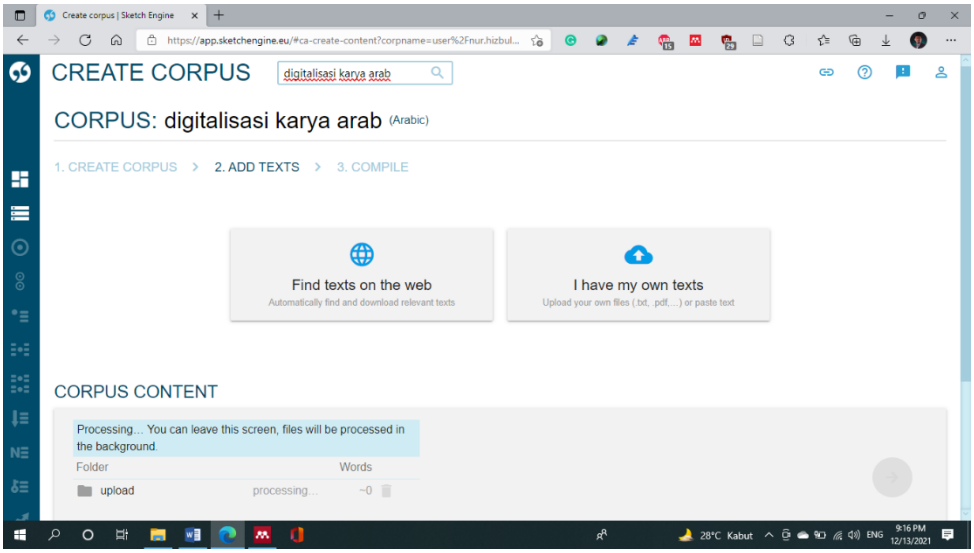
melalui mesin aplikasi korpus. Setelah data tersebut *clean*, barulah dikonversi ke format Plain Text (*txt) dengan Unicode UTF-8 seperti terlihat di gambar 3, 4 dan 5. Berikut alur proses konversi data format *doc ke format Plain Text (*txt) :

Bagan 2. Alur proses konversi data format *doc ke format Plain Text (*txt)



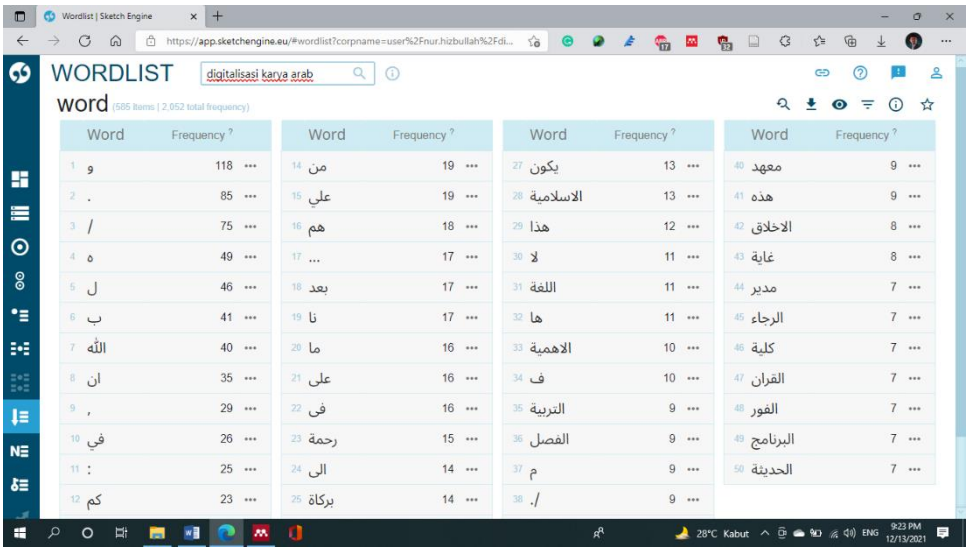
Setelah proses tersebut selesai, data tersebut telah siap untuk diolah melalui mesin aplikasi korpus, seperti *SketchEngine*. Sketchengine dipilih karena kemudahan dalam membuat file korpus secara mandiri dan dapat digunakan sesuai tujuan penelitian. Ada beberapa proses pengolahan data yang ditawarkan oleh mesin aplikasi korpus ini, diantaranya Wordlist, N-Gram dan Corcondance.[21] Berikut contoh data digital berbasis korpus yang telah diolah melalui mesin aplikasi korpus, seperti *SketchEngine* (https://app.sketchengine.eu/):

Gambar 6. Pembuataan File korpus

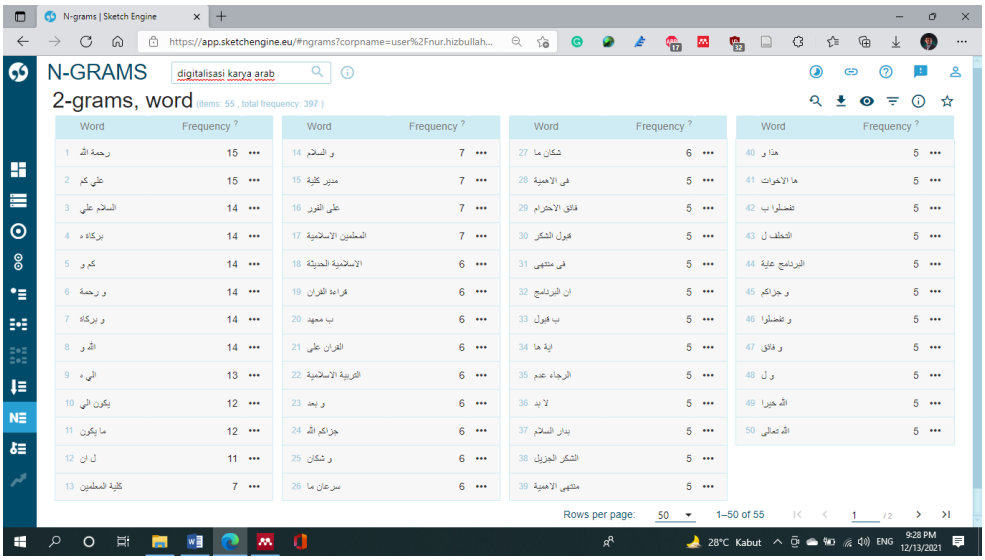


[21] Suryadarma and Fakhiroh, "Tashmīm Qāmus 'al-Zirā'Ah' Kawasīlah Ta'allum Al-'Arabiyyah Li Thalabah Qism Al-Tiknūlūjiyā Al-Shinā'iyyah Al-Zirā'iyyah Muassasn 'Alā Al-Mudawwanah Al-Lughowiyyah."

Gambar 7. Salah satu contoh fitur Wordlist



Gambar 8. Salah satu fitur N-Gram



Dari sini, dapat terlihat bahwa proses diatas dilakukan terhadap semua data konvensional yang didapatkan oleh peneliti baik melalui proses penelusuran dokumentasi, wawancara ataupun observasi langsung ke lapangan. Setelah data tersebut terkumpul kemudian diidentifkasi dan dikelompokan berdasarkan data tertulis dan data terucap seperti yang terdapat di table 1 dan 2. Penggolongan ke data calon korpus ini sesuai dengan yang diutarakan oleh Baker dalam Azzahra (2020), bahwa korpus tidak hanya berisi kumpulan teks bentuk tulis, tetapi juga mencakup ujaran. [22] Artinya sumber

[22] Azzahra, Hizbullah, and Suryaningsih, "Penyusunan Kamus Kedokteran Arab – Indonesia Dengan Pendekatan Linguistik Korpus."

data untuk data korpus dapat berbentuk tulisan dan ada juga data yang berbentuk lisan atau ucapan.

Setelah itu dimulailah proses pemindahan data (data transferring) dari data konvensioal ke data digital baik itu dilakukan secara manual maupun langsung pengubahan format dokument sebagaimana prosesnya terlihat di bagan 1 dan 2. Sedangkan proses konversi file yang sudah berbentuk format *doc yang ke format (*txt) ditunjukan dalam Gambar 3,4 dan 5. Adapun penggunaan data digital yang telah dikonversi ke dalam format *txt menggunakan *sketchengine* terlihat di gambar 6,7 dan 8. Dengan demikian proses digitalisasi karya-karya pelajar *non Arabic Speaker* di Pondok Modern Darussalam Gontor telah selesai dilakukan.

Penutup

Berdasarkan pembahasan diatas, maka peneliti menyimpulkan bahwa, pertama, karya-karya pelajar bahasa Arab *non Arabic Speaker* di Pondok Modern Darussalam Gontor (PMDG) Ponorogo Jawa Timur berdasarkan jenis datanya dapat dikategorikan menjadi dua bagian yaitu karya bahasa Arab tertulis *(written product)* dan karya bahasa Arab terucap (*Spoken Product)*, sedangkan jika dilihat dari tingkat keterbacaanya, dapat dikategorikan menjadi data konvensional tertulis yang dapat dibaca oleh software dan data konvensional yang tidak dapat dibaca. Hasil dari karya tersebut merupakan produk asli para pelajar PMDG yang telah mengkristal sedemikan rupa menjadi produk keseharian yang memang dilaksanakan secara terus menerus dan terkontrol melalui sistem disiplin pesantren.

Kedua, Proses penghimpunan dan proses digitalisasi sumber data korpus Arab karya-karya pelajar bahasa Arab *non Arabic Speaker* di Pondok Modern Darussalam Gontor Ponorogo Jawa Timur dilakukan melalui tiga tahap. Tahap pertama mengkonversi data konversional yang berupa tulisan tangan dan rekaman suara ke data digital dalam format *doc. Tahap kedua, konversi data digital dalam format *doc ke dalam format plaint text (*Txt). Tahap ketiga, memasukan data dalam format *txt ke dalam mesin olah data berbasis web yaitu *SketchEngine*. Dengan demikian data korpus digital tersebut telah siap untuk diolah menjadi suatu kajian berdasarkan tujuan kebahasaan tertentu.

Bertolak dari sini, peneliti sadar bahwa penelitian ini masih jauh dari kata sempurna diperlukan penelitian mendalam dalam skala yang lebih besar untuk mendapatkan data korpus pesantren secara lebih intensif dan maksimal. Oleh karena itu penelitian sejenis terutama mengenai korpus pesantren masih sangat layak untuk diteliti dan dianalisis lebih dalam.

Daftar Pustaka

Adolphs, Svenja. *Introducing Electronic Text Analysis A Practical Guide for Language and Literary Studies*. 1st ed. New York: Routledge, 2006.

Al-sulaiti, Latifa, Noorhan Abbas, Claire Brierley, Eric Atwell, and Ayman Alghamdi.

"Compilation of an Arabic Children's Corpus." In *LREC 2016: 10th Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari. Portorož, Slovenia, 2016. https://eprints.whiterose.ac.uk/100839/.

Alfaifi, Abdullah. "The Arabic Learner Corpus Website." Last modified 2015. https://www.arabiclearnercorpus.com/.

Alfaifi, Abdullah, and Eric Atwell. "Potential Uses of the Arabic Learner Corpus" (2013). Accessed December 14, 2021. http://www.uclouvain.be/en-cecl-longdale.html.

Azzahra, Siti Fatimah, Nur Hizbullah, and Iin Suryaningsih. "Penyusunan Kamus Kedokteran Arab – Indonesia Dengan Pendekatan Linguistik Korpus." *Tsaqofiya : Jurnal Pendidikan Bahasa dan Sastra Arab* 2, no. 2 (2020): 60–66.

Emzir. *Metodologi Penelitian Kualitatif : Analisis Data*. 6th ed. Depok: Rajawali Pres, 2018.

Hizbullah, Nur, Zaqiatul Mardiah, Yoke Suryadarma, Luthfi Muhyiddin, Oyong Sofyan, and Ferry Hidayat. "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia." *KnE Social Sciences*, no. July (2019): 3–4.

———. "Arabic Learners' Corpora in Pesantrens for Developing Arabic Language Researches in Indonesia." *KnE Social Sciences* 2019 (2019): 980–989. https://www.kne-publishing.com/index.php/KnE-Social/article/view/4922.

Hizbullah, Nur, and Muchlis Madian Muhammad. "Projected Characteristics and Content of Arabic Corpus in Indonesia." *Advances in Social Science, Education and Humanities Research (ASSEHR)* 154, no. Icclas 2017 (2018): 172–174.

Hizbullah, Nur, Fazlur Rachman, and Fuzi Fauziah. "Linguistik Korpus Dalam Kajian Dan Pembelajaran Bahasa Arab Di Indonesia." In *Konferensi Nasional Bahasa Arab (KONASBARA) II*, 385–393, 2016.

Nesselhauf, Nadja. *Corpus Linguistics : A Practical Introduction*. *Anglistisches Seminar*. Heidelberg: Uniheidelberg, 2011. http://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus-Linguistics-Practical -Introduction.pdf.

Sugiyono. *Metode Penelitian Dan Pengembangan (Research and Development)*. 4th ed. Bandung: Alfabeta, 2019.

———. *Metode Penelitian Pendidikan*. Cetakan 27. Bandung: CV. Alfabeta, 2018.

Suryadarma, Yoke, and Alinda Zakiyatul Fakhiroh. "Optimalisasi Penggunaan Corpus Linguistics Dalam Penyusunan Kamus Az- Ziro'ah Sebagai Media Pembelajaran Bahasa Arab." In *International Seminar on Language, Education, and Culture (ISoLEC) 2020*, 123–128. Malang: Universitas Negeri Malang, 2020. http://isolec.um.ac.id/proceeding/index.php/issn/article/view/59.

———. "Tashmīm Qāmus 'al-Zirā'Ah' Kawasīlah Ta'allum Al-'Arabiyyah Li Thalabah Qism Al-Tiknūlūjiyā Al-Shinā'iyyah Al-Zirā'iyyah Muassasn 'Alā Al-Mudawwanah Al-Lughowiyyah." *LISANUDHAD* 7, no. 2 (December 17, 2020): 37–56. Accessed October 20, 2021. https://ejournal.unida.gontor.ac.id/index.php/lisanu/article/view/6744.

Zaid, Abdul Hafidz. "تكنولوجيا التعليم المقترحة لتعليم مهارة الكلام لطلاب المستوى المتوسط في إندونيسيا." *LISANUDHAD* 1, no. 2 (December 8, 2014). Accessed November 7, 2020. https://ejournal.unida.gontor.ac.id/index.php/lisanu/article/view/446.