

Penyusunan Model Korpus Al-Qur'an Digital

Nur Hizbullah¹, Fazlurrahman², Fuzy Fauziah³

^{1, 2, 3} Program Studi Sastra Arab, Fakultas Sastra, Universitas Al Azhar Indonesia
Jalan Sisingamangaraja, Kebayoran Baru, Jakarta Selatan 12110

Penulis untuk Korespondensi/E-mail: nurhz@uai.ac.id

Abstrak – Penelitian ini bertujuan menyusun sebuah model file korpus Al-Qur'an digital yang dapat digunakan sebagai bahan data primer bagi penelitian kebahasaan dalam kerangka cabang ilmu linguistik korpus yang berkenaan dengan daftar kata (*word list*) dan konkordansi (*concordance*) dalam Al-Qur'an. Penelitian ini menggunakan metode kombinasi antara eksplorasi dan eksperimen yang digunakan untuk mencari berbagai aplikasi pengolah korpus dan menguji coba satu persatu aplikasi itu untuk mengolah korpus teks Al-Qur'an bertulisan Arab dengan segala karakteristiknya. Setelah aplikasi yang tepat ditemukan, langkah berikutnya menggunakan metode deskriptif yaitu menguraikan secara faktual mekanisme pengolahan bahan digital menjadi format korpus Al-Qur'an sekaligus menyusun daftar kata dan konkordansinya dalam Al-Qur'an. Penelitian ini menunjukkan bahwa aplikasi WordSmith adalah yang paling memadai untuk melakukan pengolahan teks berbahasa Arab dalam kerangka linguistik korpus. Dengan prosedur dan langkah-langkah yang sesuai dengan sistematika aplikasi tersebut, dapat dihasilkan *file* Al-Qur'an digital yang memenuhi syarat teknis untuk diolah guna menyusun daftar kata dan konkordansi.

Kata kunci : linguistik korpus, korpus Al-Qur'an, daftar kata, konkordansi

Abstract – This research aims to develop a model of the corpus file digital Qur'an that can be used as primary data for the study of language in terms of corpus linguistics branch of science with regard to word list and concordance in the Qur'an. This research is using a combination method of exploration and experimentation that are used to search for a variety of corpus processing applications and tested one by one to process Arabic Qur'an text with all its characteristics. And after having found the right application, the next step is using descriptive method that describes in factual material processing mechanisms into digital format at the same corpus of the Qur'an and its concordance and word list. This study shows that the WordSmith software is the most adequate to do the Arabic text processing within the framework of corpus linguistics. With procedures and measures in accordance with the application, it can produces a digital file of the Qur'an that is technically qualified to be processed in order to compile a list of words and a concordance.

Keywords: corpus linguistics, corpus Al-Qur'an, word list, concordance

PENDAHULUAN

K eberadaan Al-Qur'an sesungguhnya menjadi sumber utama bagi ajaran Islam. Selain itu, Al-Qur'an juga penting kedudukannya sebagai acuan utama berbagai disiplin ilmu, baik ilmu alam, ilmu pasti, maupun ilmu sosial-humaniora. Bagi salah satu cabang dalam rumpun ilmu humaniora, yaitu

ilmu bahasa, dalam hal ini bahasa Arab, Al-Qur'an merupakan sumber data primer karena memiliki ragam bahasa yang baku dan tinggi, baik dalam aspek gramatika, retorika, stilistika, maupun sisi kebahasaan Arab yang lain.

Kajian kebahasaan terhadap Al-Qur'an sudah menjadi perhatian para linguist Arab maupun non-Arab sejak dahulu kala. Orientasi kajian pun beragam, mulai dari tataran bahasa paling

rendah, seperti fonem, morfem, lalu setingkat di atasnya, seperti kajian kosa kata, sampai ke tataran yang lebih tinggi dan luas, yaitu kajian struktur kalimat, kajian makna, dan kajian wacana. Beragam kajian itu memposisikan Al-Qur'an sebagai sebuah teks, dalam hal ini teks kebahasaan. Namun, model teks Al-Qur'an sendiri berkembang dari zaman ke zaman. Sebelum era digital, teks Al-Qur'an hanya dituliskan dengan tangan di atas kertas kuno atau tertera di atas lembaran kertas yang dicetak secara konvensional lazimnya beredar di kalangan umat Islam dan bangsa lain. Di era modern, khususnya di era digital, teks Al-Qur'an telah dikonversi ke banyak model teks dan bisa diolah melalui berbagai piranti lunak maupun aplikasi digital khusus, sekali lagi, dalam konteks Al-Qur'an sebagai sebuah teks kebahasaan.

Situasi itu kemudian memposisikan Al-Qur'an sebagai "korpus". Dalam ilmu linguistik, korpus adalah sekumpulan data, bisa konvensional maupun digital, dalam bentuk tertulis yang mengandung berbagai informasi kebahasaan untuk diteliti mulai dari tataran kata, struktur, makna, dan wacana. Dalam perkembangan terkini, teks Al-Qur'an banyak diolah dalam bentuk digital dan menjadi bahan utama yang sangat penting bagi kajian linguistik terhadap kitab suci pada era penelitian lebih lanjut.

Di masa kini, berkembang suatu cabang ilmu di bawah rumpun ilmu linguistik, yaitu linguistik korpus. Menurut pengertiannya, linguistik korpus meneliti bahasa melalui seperangkat data yang sifatnya alamiah, apa adanya, baik itu data tulisan maupun data lisan yang ditranskripsikan atau dituliskan (Adolphs, 2006: 137). Menurut Cheng (2012: 6), linguistik korpus menganalisis data dengan mengompilaskannya terlebih dahulu berdasarkan kategori dan satuan analisisnya. Lebih lanjut, kategori dan satuan analisis yang paling penting dalam linguistik korpus adalah daftar kata (*word list*) dan konkordansi (*concordance*). Daftar kata berisi semua kata yang terdapat dalam data dan disertai dengan jumlah frekuensi kemunculan kata dalam teks. Adapun konkordansi adalah daftar kata-kata yang berkolokasi dan membentuk suatu makna berdasarkan konteks dan koteksnya (Adolphs, 2006: 136). Dahulu kala, semua data kebahasaan tulisan masih berupa manuskrip

atau tulisan konvensional. Sejak ditemukannya teknologi komputer dan dimulainya era digital, kini data kebahasaan berupa tulisan atau teks serta bahasa lisan yang dituliskan dapat disimpan dan diolah secara digital dengan lebih mudah. Di sinilah linguistik korpus mengambil bagian, yaitu mengolah data kebahasaan yang sudah berbentuk digital.

Apabila dulu Al-Qur'an hanya berupa mushaf konvensional bertulisan tangan dan lalu dicetak dengan berbagai bentuk dan variasi tulisan, seiring dengan perkembangan dan kemajuan teknologi, kini Al-Qur'an pun memasuki era digital. Format digital Al-Qur'an kini tersedia dalam berbagai versi: ada yang berupa gambar dalam aplikasi Al-Qur'an digital, ada yang berupa *font* yang dapat diinstalasikan ke dalam sistem operasi komputer, maupun teks digital biasa yang bisa diakses dan diolah dengan aplikasi pengolah kata. Berbagai bentuk teks digital yang sudah ada setakat ini cukup bisa dimanfaatkan dalam beberapa penggunaan yang terbatas, seperti aplikasi Al-Qur'an digital untuk bacaan, dan lain sebagainya. Namun, untuk penelitian yang memanfaatkan format digital teks Al-Qur'an, format yang ada masih harus diolah dan diverifikasi agar dapat dijadikan bahan dasar bagi format baru yang disebut dengan korpus Al-Qur'an.

Untuk menyusun format tersebut, penelitian ini akan melakukan sejumlah langkah guna menyiapkan suatu model data yang nantinya dapat dimanfaatkan untuk berbagai kajian kebahasaan terhadap Al-Qur'an oleh berbagai pihak yang berminat kepada kajian bahasa Arab Al-Qur'an, terutama yang menggunakan linguistik korpus. Diharapkan agar produk atau hasil kajian itu nantinya dapat membantu mempermudah masyarakat dalam memahami Al-Qur'an serta dan menyingkap banyak sisi kebahasaan yang kaya akan informasi penting seputar ajaran dan nilai Islam.

Dengan latar belakang situasi tersebut, penelitian ini sesungguhnya bertujuan menyusun sebuah model file korpus Al-Qur'an digital yang dapat digunakan sebagai bahan data primer bagi penelitian kebahasaan dalam kerangka cabang ilmu linguistik korpus yang berkenaan dengan daftar kata (*word list*) dan konkordansi (*concordance*) dalam Al-Qur'an.

Dengan mengacu kepada tujuan penelitian tersebut, maka rumusan masalah penelitian ini terfokus pada model file korpus Al-Qur'an digital yang dapat dijadikan bahan data primer bagi penelitian kebahasaan dalam kerangka linguistik korpus. Adapun pertanyaan penelitian yang diajukan adalah:

1. Aplikasi apa yang paling representatif digunakan untuk membuat model file korpus Al-Qur'an digital?
2. Bagaimana format file yang dapat diolah aplikasi tersebut?
3. Bagaimana format isi teks digital yang akan diolah dengan aplikasi tersebut?
4. Bagaimana teknik membuat daftar kata dan konkordansi dalam Al-Qur'an dengan menggunakan aplikasi tersebut?

Bidang penelitian linguistik korpus sesungguhnya sangatlah luas. Demikian pula dengan Al-Qur'an sebagai objek penelitian. Oleh karena itu, lingkup penelitian ini dibatasi hanya pada aspek pengolahan bahan data digital bernama "*Al-Mushaf Al-Imla'i*" dengan menggunakan aplikasi pengolah korpus yang representatif untuk menghasilkan format digital standar yang dapat digunakan untuk melakukan pemetaan daftar kata dan konkordansi dalam Al-Qur'an.

Penelitian ini diharapkan nantinya secara khusus dapat memberikan kontribusi berupa :

1. Terbentuknya model teks digital Al-Qur'an yang bisa digunakan untuk pemetaan daftar kata dan konkordansi;
2. contoh model pemetaan daftar dan konkordansi kata dan kalimat dalam Al-Qur'an, serta
3. gambaran langkah, prosedur, dan sistematika penelitian korpus Al-Qur'an menggunakan bahan digital.

Penelitian ini memang hanya terfokus pada Al-Quran digital sebagai model awal. Namun demikian, nantinya model dan prosedur yang ada dapat digunakan secara luas dalam berbagai kajian maupun keperluan di bidang bahasa Arab. Sebagai contoh, penyusunan daftar kata ataupun kamus bahasa di bidang khusus dapat menggunakan pendekatan ini. Daftar istilah bahasa Arab di bidang diplomasi, misalnya, dapat diambilkan dari sejumlah teks berbahasa Arab yang berkisar pada topik-topik kediplomasian. Manfaat berikutnya, daftar istilah tersebut dapat digunakan sebagai salah

satu bahan ajar bahasa Arab diplomasi sebagai kosa kata yang harus dikenali dan dikuasai oleh mahasiswa/pembelajar. Demikian pula di bidang-bidang lainnya, penelitian ini nantinya dapat menawarkan suatu metodologi penyusunan korpus dari suatu teks sumber kebahasaan.

TEORI TENTANG LINGUISTIK KORPUS

Penelitian ini sesungguhnya adalah sebuah model dari kajian dalam bidang ilmu linguistik korpus. Menurut pengertiannya, linguistik korpus adalah sebuah metode empiris dalam analisis dan deskripsi linguistik untuk meneliti bahasa yang dipakai oleh para penutur secara alamiah. Bahasa itu dikompilasikan berdasarkan kategori tertentu lalu diteliti dengan tujuan mencapai pemahaman yang benar dari pemakaian bahasa tersebut (Cheng, 2012:6, 29-30). Pada tahun 1980-an, Linguistik Korpus mulai me-rambah kepada pemanfaatan teknologi komputer dan itu cukup membangkitkan gairah perkembangan metodologi korpus. Sejak itu, kemunculan sejumlah korpus dan kajian berbasis korpus semakin meningkat drastis. Pada masa kini, metodologi korpus sudah sangat populer dan bahkan digunakan di berbagai macam riset multidisiplin ilmu pengetahuan.

Dalam hal penyusunan korpus bahasa, kini sudah tersedia banyak sekali model korpus bahasa. Wikipedia, misalnya, mendata ada sejumlah korpus bahasa yang dibagi ke dalam beberapa kelompok bahasa. Uniknya, bahasa Inggris berada dalam kelompok tersendiri karena ketersediaan banyak sekali ragam dan sumber korpusnya, seperti yang paling terkemuka dan termutakhirkan adalah British National Corpus (BNC), American National Corpus (ANC), Corpus of Contemporary American English (COCA), dan lain sebagainya. Selain bahasa Inggris korpus bahasa-bahasa dunia dimuat daftarnya di laman wikipedia dalam kelompok-kelompok bahasa Eropa, bahasa Timur Tengah, bahasa Asia Timur, korpus paralel multibahasa, dan korpus perbandingan antarbahasa.¹

¹

https://en.wikipedia.org/wiki/List_of_text_corpora, diakses 29 Juli 2016

Dalam penyusunan korpus bahasa Arab, melalui aplikasi Sketch Engine, misalnya, telah tersedia korpus yang berisi lebih kurang 5,8 juta kata (versi 2012) dan dikembangkan menjadi 8,3 juta kata (versi 2015)² yang diambil dari berbagai sumber. Selain itu, tercatat ada sekitar 18 jenis korpus dari berbagai sumber dalam sublam University of Leeds, di antaranya Corpus of Contemporary Arabic, Arabic Gigaword, dan yang sedang dalam pengembangan adalah International Corpus of Arabic oleh University of Alexandria, Mesir.³

Di samping sudah tersedianya korpus bahasa Arab, disusun pula korpus khusus mengenai Al-Qur'an dengan menggunakan berbagai pendekatan dan aplikasi. Yang paling representatif kiranya adalah karya University of Leed, yaitu *The Quranic Arabic Corpus*.⁴ Lebih dari sekadar korpus Al-Qur'an, laman itu menyediakan berbagai peta analisis kebahasaan Al-Qur'an, seperti terjemah kata-per-kata dengan bahasa Inggris, kamus Al-Qur'an, "pohon" sintaksis struktur konsep dan kata dalam Al-Qur'an, dan gramatika Al-Qur'an. Laman ini menyediakan pula file Al-Qur'an digital dalam format *.txt yang bisa diunduh untuk bahan analisis. Namun, file tersebut hanya berisi teks Al-Qur'an bertransliterasi latin, bukan Arab, meski dilengkapi dengan status gramatikal setiap kata dalam Al-Qur'an. Hal itu menjadi tantangan tersendiri bagi langkah analisis teks Al-Qur'an dengan aplikasi pengolah korpus.

Pendalaman lebih lanjut dilakukan Sharaf (2009). Melalui situs resminya, Sharaf memperkenalkan serangkaian kajiannya berikut hasil pemetaan konseptual dengan aplikasi khusus. Namun, tidak dapat ditemukan bahan digital olahannya yang dapat diunduh dan dimanfaatkan oleh para peneliti lain untuk melakukan ataupun melanjutkan pekerjaannya.⁵

Selain laman web, ada pula beberapa aplikasi khusus yang berisi data dan analisis gramatika

kebahasaan Al-Qur'an, di antaranya adalah QuranCode. Namun, aplikasi ini hanya memberikan gambaran jumlah kata dan gramatika kata-kata dalam Al-Qur'an meski dilengkapi dengan informasi klasifikasi ayat berdasarkan kategori makiyah dan madaniyah, pencarian kata, dan lain-lain.⁶ Aplikasi ini belum optimal fungsinya jika harus digunakan untuk keperluan analisis korpus yang lebih luas.

Yang menjadi catatan penelitian ini sampai bagian ini adalah belum adanya model file Al-Qur'an digital yang tersedia dan dapat digunakan untuk melakukan analisis korpus serta cara pengolahan file tersebut dengan aplikasi tertentu. Atas dasar itulah, penelitian ini memandang perlunya penyusunan model korpus Al-Qur'an digital untuk penelitian dengan pendekatan linguistik korpus.

Penelitian ini mengacu antara lain kepada sejumlah penelitian terdahulu yang terkait dengan korpus bahasa Arab dan khususnya korpus Al-Qur'an. Di Indonesia, kajian linguistik korpus tergolong baru, apalagi terhadap teks-teks keagamaan seperti Al-Qur'an. Oleh karena itu, dipilihlah beberapa penelitian dari luar negeri yang berkenaan dengan korpus Al-Qur'an berikut ini.

Khan dan Alginahi (2013) menulis laporan penelitian tentang tantangan dan keperluan digitalisasi Al-Qur'an. Dengan metode survei menggunakan kuesioner, mereka mencoba melihat sejauh mana kebutuhan umat Islam terhadap upaya digitalisasi Al-Qur'an dalam konteks penggunaannya untuk keperluan rutin seperti membaca dan mempelajarinya. Para peneliti menyimpulkan, secara umum, umat Islam yang menjadi responden penelitian mereka, terlebih generasi muda yang sudah akrab dengan teknologi, merasa perlunya langkah segera dan menyeluruh bagi digitalisasi Al-Qur'an. Untuk mendukung hal itu, penelitian mereka merekomendasikan agar format digital Al-Qur'an dibuat semudah dan seramah mungkin dengan pengguna serta diawasi dan dikendalikan oleh lembaga yang berwenang agar tidak terjadi dan tersebar kesalahan-kesalahan yang dapat merusak kesucian Al-Qur'an.

² <https://www.sketchengine.co.uk/artenten-corpus/> diakses 29 Juli 2016

³ http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpus.htm, diakses 29 Juli 2016

⁴ <http://corpus.quran.com/> diakses 29 Juli 2016

⁵ <http://www.textminingthequran.com/wiki/QurCord>, diakses 29 Juli 2016

⁶ <http://heliwave.com>, diakses 29 Juli 2016

Salah satu langkah penyusunan korpus bahasa Arab yang mencantumkan Al-Qur'an sebagai salah satu sumbernya adalah yang dilakukan oleh Alrabia (et.al., 2014). Mereka menamakan korpusnya dengan King Saud University Corpus of Classical Arabic. Sesuai namanya, mereka hanya mengambil korpus dari bahasa Arab periode klasik dan Al-Qur'an termasuk ke dalam periode itu. Mereka beralasan bahwa pola bahasa Arab klasik adalah basis teori linguistik Arab dan harus diperlakukan serta dipahami secara akademis. Korpus ini mencakup 50 juta kata dari khazanah bahasa Arab klasik dan digunakan untuk kepentingan studi distribusi semantik leksikal Al-Qur'an dan bahasa Arab klasik dengan pendekatan linguistik komputasional. Namun, selain menggambarkan struktur korpus, penelitian mereka juga mendeskripsikan gejala perubahan bahasa dari masa klasik ke masa kini.

Sharaf dan Atwell (2012). Mereka membuat "QurAna", korpus Al-Qur'an beranotasi pada anafora pronominal. Penelitian mereka hanya terfokus pada analisis pronomina dalam Al-Qur'an dan pola kemunculan anafora yang terkait dengan pronomina dan antesedennya dalam teks-teks ayat Al-Qur'an. Model analisis penelitian ini berguna bagi analisis pronomina, anafora, dan antesedennya dalam wacana bahasa Arab modern dan perbandingan antara kedua masa itu.

Salah satu contoh lain penelitian korpus Al-Qur'an namun lebih spesifik adalah karya Imad dan Abdelhak (2016) melakukan penelitian untuk menyusun korpus Al-Qur'an khusus untuk penelitian morfologi Arab. Mereka menggunakan teks Al-Qur'an dan tabel morfologi Al-Khalil. Mereka menggunakan Quranic Arabic Corpus milik University of Leeds dan Quranic Corpus of Haifa milik University of Haifa sebagai pedoman dan lalu membuat analisis dan kritik terhadap konsep dasar akar kata, serta diakhiri dengan perbandingan antara keduanya dan pola rumusan si peneliti sendiri. Dalam pengolahan data, mereka menggunakan teknik anotasi semiotomatis dan rumusan CSV (Comma Separated Values) yang cukup kompleks. Hasilnya, rumusan pola dasar konsep morfologi mereka lebih kompleks bila dibandingkan dengan dua model acuannya dan menghasilkan lebih banyak lema atau entri kata dari keseluruhan kata dalam isi Al-Qur'an.

Berdasarkan informasi dari sejumlah penelitian terdahulu, dapat dikatakan bahwa belum ada atau belum ditemukan karya yang khusus mengkaji teknik pembuatan model file digital Al-Qur'an untuk keperluan penelitian linguistik korpus. Dari segi hasil pun, peneliti melihat belum ada model korpus Al-Qur'an yang sederhana dan dapat dijadikan model bagi pembuatan korpus, dalam hal ini, yang berbahasa Arab. Oleh karena itu, yang berbeda dari hasil penelitian ini bila dibandingkan dengan hasil penelitian terdahulu adalah pada model korpus yang dihasilkan. Selain hasil, penelitian ini juga memberikan langkah-langkah aplikatif tentang cara pembuatan model *file* korpus terutama terkait data yang berbahasa Arab.

METODE PENELITIAN

Penelitian ini menggunakan kombinasi beberapa metode terkait dengan kondisi data dan tujuan penelitian. Metode pertama adalah kombinasi antara eksplorasi dan eksperimen yang digunakan untuk mencari berbagai aplikasi pengolah korpus dan menguji coba satu persatu aplikasi itu untuk mengolah korpus teks Al-Qur'an bertulisan Arab dengan segala karakteristiknya. Setelah aplikasi yang tepat ditemukan, langkah berikutnya menggunakan metode deskriptif yaitu menguraikan secara faktual mekanisme pengolahan bahan digital menjadi format korpus Al-Qur'an sekaligus menyusun daftar kata dan konkordansinya dalam Al-Qur'an.

Berikut digambarkan prosedur dan mekanisme kerja dari penelitian ini menurut tahapan-tahapannya.

- a. Pemerolehan teks digital Al-Qur'an dan konversinya. Teks dalam bentuk dokumen (*.doc) yang telah diunduh diformat ulang ke dalam bentuk teks (*.txt) dengan konversi model UTF-8 pada proses *save-as*-nya di MS Word.
- b. Bahan korpus diverifikasi dengan cara dibaca secara manual dan otomatis untuk memastikan kelengkapan ayat Al-Qur'an secara keseluruhan dan meng-hindari jangan sampai ada ayat yang hilang akibat proses konversi digital.
- c. Penelusuran, percobaan, dan pemilihan aplikasi pengolah korpus yang memungkinkan proses pengolahan teks

- berbahasa Arab dan penerapan fungsi penyusunan daftar kata dan konkordansi.
- d. Dari bahan korpus yang sudah diverifikasi, dibuatlah contoh model penyusunan daftar kata dan konkordansi kata dalam Al-Qur'an secara digital.

Data penelitian ini berupa ayat-ayat Al-Qur'an yang terdiri atas kata dan kalimat yang bersumber dari teks digital “*Al-Mushaf Al-Imla'iy*”. Bahan tersebut diunduh dari tanzil.net.⁷ Karakteristik teks ini di antaranya berasal dari teks periwayatan termasyhur di kalangan para ahli qiraah, yaitu riwayat Hafs dari ‘Ashim. Teks ini juga mengacu kepada mushaf dengan ortografi Usmani, yaitu mushaf yang merujuk kepada kodifikasi Usman bin ‘Affan. Lebih lanjut, karena ada perbedaan ortografi dengan model ortografi Arab standar modern, maka ortografi Usmani itu disesuaikan menjadi ortografi Arab standar modern. Itulah yang disebut dengan *Al-Mushaf Al-Imla'iy*.

HASIL DAN PEMBAHASAN

Pemerolehan File Al-Qur'an Digital Mushaf Imla'i dan Pengoversiannya

Ismail dan Rahman (2016: 44) menyebutkan, *mushaf Imla'i* merupakan versi mushaf Al-Quran yang ditulis berdasarkan kaidah penulisan bahasa Arab biasa atau standar baku yang disebut *Rasm Imla'i* atau *Rasm Qiyasi*. Struktur penulisan kata dalam Al-Quran (*rasm*) dan tanda-tanda (*dabt* atau diakritik) dalam *Mushaf Imla'i* memiliki perbedaan dari apa yang ada dalam *Mushaf Usmani*, yaitu mushaf standar yang ditulis sejak era Khalifah Usman bin ‘Affan. Secara khusus, perbedaan tersebut dapat dilihat pada lima faktor teknik ortografi, yaitu pembuangan huruf (*hadhf*), penambahan huruf (*ziyadah*), penukaran huruf (*badl*), penyambungan dua kata (*wasl*), dan pemisahan dua kata (*fasl*).

Faktor	<i>Mushaf Imla'i</i>	<i>Mushaf Usmani</i>
Pembuangan huruf	خَالِدُونَ	خَالِدُونَ
Penambahan huruf	يَمُوسَى	يَايُوسَى
Penukaran huruf	يَسْطُ	يَسْطُط
Penyambungan dua kalimat	يَا مُوسَى	يَا مُوسَى
Pengasingan dua kalimat	إِذَا	إِذَا

Gambar 1. Contoh perbedaan *Mushaf Imla'i* dari *Mushaf Usmani* dalam penelitian Ismail dan Rahman (2016: 44)

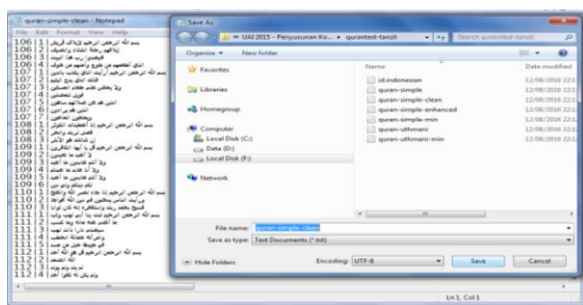
Secara digital, kedua jenis tulisan *Mushaf Usmani* dan *Mushaf Imla'i* sudah ada dan tersebar luas penggunaannya dalam berbagai aplikasi. Namun, untuk aplikasi digital pengolahan kata, *Mushaf Imla'i* lebih banyak dipakai karena bentuknya yang standar dengan kaidah penulisan Arab standar baku hingga kini. Adapun *Mushaf Usmani* tidak banyak digunakan karena kekhasan karakteristik tulisannya dan dianggap tulisan baku Al-Qur'an sejak masa awal penulisannya dan bahkan ada sebagian ahli ortografi Arab yang memperlakukannya sebagai tulisan yang “sakral” dan tak boleh diganggu-gugat apalagi dimodifikasi menjadi format lainnya.

Untuk keperluan pengolahan teks digital, penelitian ini memanfaatkan teks yang tersedia pada laman tanzil.net.⁸ Teks yang tersedia ada dua jenis, yaitu *Usmani* dan *Imla'i*. Namun, yang dimanfaatkan untuk penelitian ini adalah yang berjenis *Imla'i*. Lebih lanjut, teks yang dipilih tidak mengandung tanda-tanda diakritik, jadi hanya berisi rangkaian huruf dalam semua kata. Dengan demikian, pengolahan file dalam aplikasi korpus akan lebih mudah dan terbebas dari kendala teknis.

Setelah file tersebut didapatkan, perlu dipastikan bahwa *encoding* konversinya menggunakan UTF-8 dan disimpan dengan ekstensi *.txt dari semula *.doc. Pemilihan dan penggunaan *encoding* UTF-8 disarankan dalam petunjuk teknis penggunaan dan pengolahan file di laman tanzil.net. Setelah diuji coba, dapat dikonfirmasi bahwa mode konversi selain *encoding* UTF-8 tak berterima bagi hampir semua aplikasi pengolah korpus. Secara teknis, hanya file dengan ekstensi *.txt dengan mode konversi UTF-8 saja yang dapat diolah dengan aplikasi korpus.

⁷ <http://tanzil.net/>

⁸ <http://tanzil.net/download>



Gambar 2. Mekanisme konversi *file* Al-Qur'an digital dari *.doc ke *.txt dengan *encoding* menggunakan UTF-8 dan contoh teks *Mushaf Imla*'i

Verifikasi Bahan Korpus

Sebelum diolah dengan aplikasi, bahan korpus yang sudah dikonversi kemudian harus diverifikasi dengan cara dibaca secara manual untuk memastikan kelengkapan ayat Al-Qur'an secara keseluruhan dan menghindari jangan sampai ada ayat yang hilang akibat proses konversi digital. Hal ini sengaja dilakukan secara khusus terhadap teks Al-Qur'an karena statusnya sebagai kitab suci demi menjaga agar jangan sampai ada unsur teks yang hilang karena pasti berdampak kepada hasil akhir dari proses pengolahan korpus. Bagaimana pun, proses konversi teks dari suatu format ke format lain selalu berisiko terhadap kondisi teks, terlebih jika teks itu ditulis dalam bahasa dengan sistem ortografi yang tidak sama dengan sistem ortografi bahasa latin.

Cara verifikasi manual yang dilakukan adalah dengan membaca teks ayat demi ayat, sambil merujuknya ke bacaan tilawah Al-Qur'an dengan tempo yang lambat. Pembacaan dilakukan secara bertahap surah demi surah, juz demi juz, sampai tuntas sebanyak 114 surah dalam 30 juz. Penting untuk dipastikan ada atau tidak unsur teks yang hilang dalam *file* digital Al-Qur'an yang akan diolah. Setelah dipastikan kelengkapan dan keutuhan teks, barulah *file* tersebut bisa diolah melalui aplikasi khusus.

Pemilihan Aplikasi Pengolah Korpus

Setelah *file* digital didapat dan dikonversi, langkah berikutnya adalah pencarian, uji coba, dan pemilihan aplikasi pengolah korpus. Aplikasi itu sesungguhnya cukup banyak yang bisa dipilih. Namun, karakteristik bahasa Arab dan ortografinya yang relatif berbeda dari banyak bahasa pada umumnya, seperti arah tulisan dari kanan ke kiri, keberadaan tanda diakritik, dan jenis serta bentuk huruf yang spesifik, membuat pilihan menjadi terbatas.

Beberapa aplikasi yang ada, misalnya, hanya mampu membaca teks Arab, tetapi tidak mampu membuat susunan kalimat dengan pola yang berlaku pada tulisan Arab (dari kanan ke kiri). Untuk memecahkan masalah tersebut, selain mencari dan mengonsultasikan masalah ini ke praktisi dan ahli linguistik korpus dari Universitas Indonesia, peneliti juga mengeksplorasi dan mempelajari beberapa pola korpus bahasa Arab yang ada dan model-model analisisnya lalu mencoba melakukan eksperimen pengolahan korpus dengan aplikasi tersebut. Berikut ulasan tentang beberapa aplikasi yang ditemukan dan dicoba untuk mengolah korpus teks Al-Qur'an.

Nooj

Nooj adalah sebuah aplikasi pengolah korpus yang memungkinkan ahli bahasa untuk melakukan analisis linguistik terhadap ortografi dan ejaan, kamus suatu topik tertentu, pola-pola kalimat, dan sebagainya. Nooj memfasilitasi penggunaanya untuk dapat memberikan deskripsi setiap fenomena linguistik serta alat yang digunakan untuk penentuan klasifikasi kelas kata dalam kalimat suatu bahasa yang dirancang dengan sangat efisien. Aplikasi ini juga memungkinkan penggunaanya untuk kuantifikasi teks secara otomatis, menyusun konkordansi, menyortir lema untuk kamus, melakukan analisis statistik, dan lain-lain.

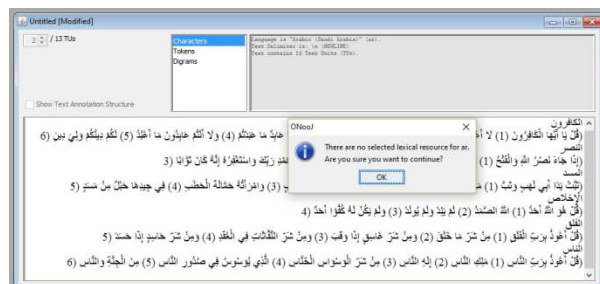
Nooj tersedia dan dapat diunduh secara cuma-cuma dengan beberapa modul untuk multibahasa, antara lain Acadian, Arab, Armenia, Bulgaria, Catalan, Cina, Kroasia, Prancis, Inggris, Jerman, Ibrani, Yunani, Hungaria, Italia, Polandia, Portugis, Spanyol dan Turki. Nooj dapat bekerja dalam sistem operasi MS-Windows, Mac OS X, LINUX dan BSD Unix.

Saat dicoba digunakan untuk analisis teks Al-Qur'an, aplikasi ini menunjukkan kekurangan dan itu menjadi kendala teknis yang tidak memungkinkan bagi analisis linguistik korpus lebih lanjut, yaitu:

- Program ini selalu memerlukan aplikasi penunjang Java untuk menjalankannya.
- File* eksekutor Nooj.exe tidak terdapat dalam *folder* program yang sudah terinstalasi, sehingga seringkali program sulit ataupun bermasalah dalam eksekusinya.

- c. Program ini tidak menunjang untuk kalkulasi daftar kata dan kolokasi, hanya bisa untuk konkordansi saja.
- d. Program ini tidak memiliki material *lexical resources* untuk bahasa Arab sehingga tidak bisa dipakai untuk menganalisis data berbahasa Arab yang kompleks dan beragam.

Berikut contoh halaman aplikasi Nooj untuk mengolah teks Al-Qur'an juz 30.

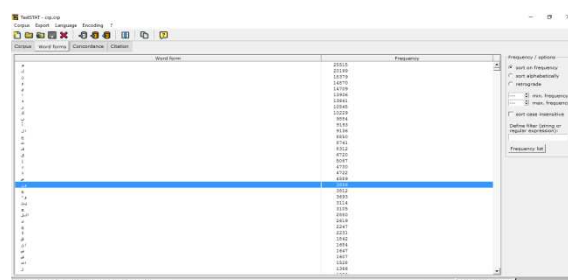


Gambar 3. Contoh halaman aplikasi Nooj

TextSTAT

TextStat adalah program yang terhitung sederhana dan mudah digunakan untuk menganalisis teks. Program ini dapat membaca berbagai macam teks dalam banyak model *encoding* dan *file* dalam bentuk HTML. Dengan program ini, pengguna dapat mengetahui frekuensi kemunculan kata dan konkordansi dari setiap file yang dibuka. TextStat dapat membaca file MS Word maupun OpenOffice tanpa harus dikonversi. Selain itu program ini dapat digunakan untuk membuka file dengan berbagai macam bahasa dan mode *encoding*. Program ini dapat di jalankan dalam system operasi windows, linux, maupun MacOS.

Namun, saat dicoba untuk mengolah teks Al-Qur'an, program ini menampilkan dalam menu *word frequency* dan *word form* masih banyak kata-kata yang terpenggal, seperti: kata الذين /alladzi:na/, kata يَنْ menjadi 1 penggalan kata sendiri, terpisah dari bagian depannya ال. Hal ini menandakan ada masalah sistemik dalam aplikasi itu untuk berinteraksi dengan teks berbahasa Arab. Amat riskan kiranya jika aplikasi itu digunakan untuk analisis teks Al-Qur'an karena pasti akan menimbulkan banyak masalah pada bentuk-bentuk kata yang terpotong dan tidak bisa diidentifikasi lebih lanjut untuk analisis linguistik.

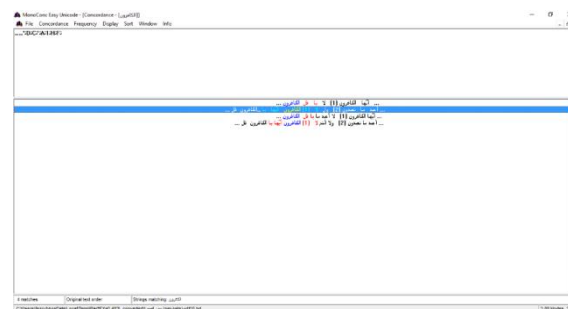


Gambar 4. Contoh halaman aplikasi TextStat

MonoconcEsy

MonoconcEsy adalah program yang didesain untuk melakukan pencarian kata ataupun ungkapan kebahasaan. Kelebihan program ini dapat mencari kata berikut dengan perubahan bentuk grammatikanya, contoh: speak – speaks – spoke – speaking, dll. Program ini merupakan program yang dapat dijalankan dalam windows 95 hingga versi yang terbaru windows 7/8/10.

Tidak jauh berbeda dari dua aplikasi sebelumnya, aplikasi ini tidak bisa digunakan untuk menganalisis teks berbahasa Arab dengan baik sebagaimana yang diperlukan untuk analisis korpus teks Al-Qur'an. Itu karena aplikasi ini hanya memiliki fitur yang terbatas, hanya dapat membaca file dengan format (.txt), tidak bisa membaca *file* selain itu, dan dalam fitur konkordansi, hasil pencarian kata yang ditampilkan terbalik-balik, sehingga menyulitkan pengguna untuk membaca hasil olahan teks.



Gambar 5. Contoh halaman aplikasi MonoconcEsy

AntConc

Seperti aplikasi yang lainnya, AntConc merupakan aplikasi yang digunakan untuk melakukan analisis teks dan konkordansi. Namun, dibandingkan dengan yang lainnya, aplikasi ini relatif lebih berkembang dan kaya menu untuk mengolah teks. Artinya, aplikasi ini mampu melakukan analisis dan penyusunan daftar kata, dan konkordansi. Lebih dari itu, menu-menu statistik pada aplikasi ini sangat

membantu dan memperkaya analisis linguistik sampai ke aspek kuantitatif yang lebih luas. Aplikasi ini bahkan bisa mengolah *file* dalam ukuran yang besar dan berisi jumlah kata yang lebih banyak. Tidak hanya itu, aplikasi ini dapat berjalan dalam sistem operasi Windows, Linux, maupun MasOS.

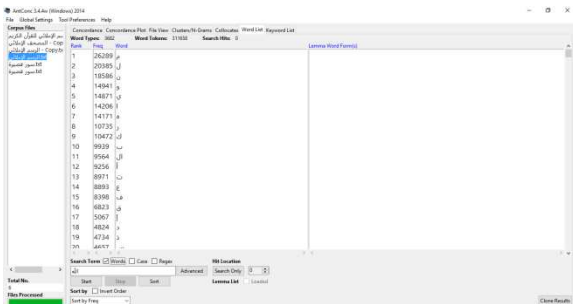
Meski secara umum lebih unggul dibanding aplikasi pengolah korpus lainnya, program ini tetap saja memiliki kendala untuk mengolah teks berbahasa Arab, yaitu:

- seringkali gagal membaca file;
- program ini hanya dapat membaca file dengan format *.txt, maka pengguna harus mengkonversi *file* dengan format yang berbeda agar dapat dibaca;
- program ini tidak dapat konsisten membaca teks berbahasa arab dengan sempurna meski *encoding* sudah menggunakan UTF-8.



Gambar 6. Contoh halaman aplikasi AntCont untuk melihat *file* konversian ke *.txt

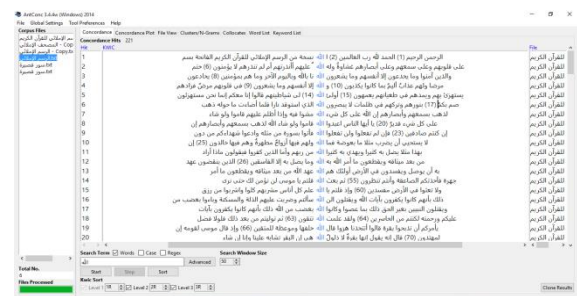
- Dalam fitur *word list*, kata-kata terpecah menjadi huruf sehingga fitur ini tidak maksimal fungsinya untuk menghitung jumlah kata dalam sebuah *file*.



Gambar 7. Contoh halaman aplikasi AntCont untuk melihat hasil pengolahan *word list*

- Dalam fitur *concordance*, susunan kata dalam hasil pencarian tidak teratur. Artinya, aplikasi AntCont gagal membaca teks berbahasa Arab dengan sistem arah tulisan dari kanan-ke-kiri. Lebih dari itu, susunan

kalimat dalam teks menjadi tidak teratur sebagaimana aslinya dan itu tidak memungkinkan untuk melakukan proses analisis konkordansi.

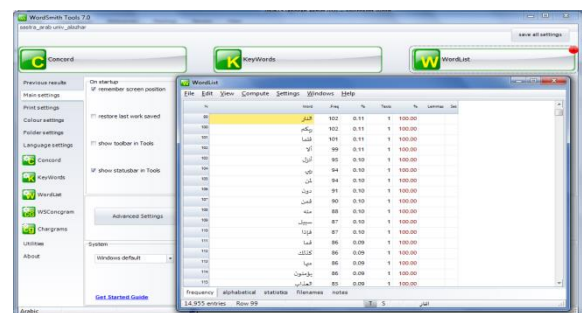


Gambar 8. Contoh halaman aplikasi AntCont untuk melihat hasil pengolahan konkordansi

WordSmith

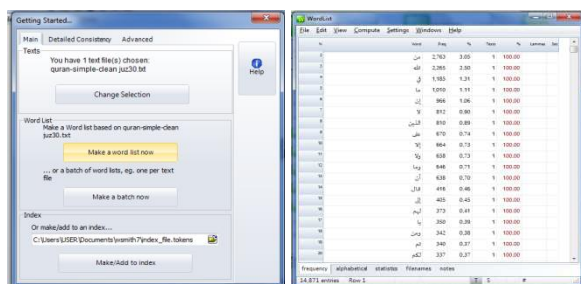
Dibanding aplikasi-aplikasi lainnya, WordSmith adalah yang paling lengkap dan paling representatif. Yang lebih utama, kemampuannya untuk mengolah teks multibahasa sudah memadai, lengkap dengan konten penyesuaian sistem yang diperlukan untuk menganalisis teks berbagai bahasa yang memiliki keunikan ortografi dan kesulitan teknis lainnya. Lebih lanjut, aplikasi ini menyediakan fasilitas konversi teks yang memungkinkan kustomisasi yang optimal bagi teks yang akan dianalisis.

Selain daftar kata dan konkordansi, aplikasi ini juga menyediakan fasilitas untuk menganalisis kata-kata kunci dalam suatu wacana yang terkandung dalam teks. Kata kunci adalah sejumlah kata yang paling produktif muncul dalam suatu teks dan diasumsikan sebagai “pusat wacana” dari teks tersebut. Namun, kata kunci bukanlah kata tugas, seperti preposisi, kata sambung dll yang memang sering muncul dalam wacana.



Gambar 9. Contoh halaman aplikasi WordSmith untuk melihat hasil pengolahan daftar kata berikut frekuensi kemunculannya

- Setelah teks yang dipilih muncul di halaman semula, klik menu *Make a word list now*, maka akan muncul halaman seperti di kanan bawah berikut.

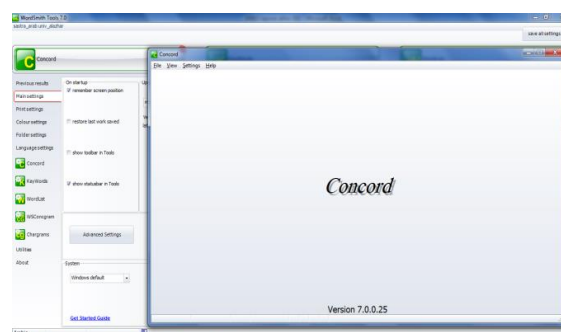


Gambar 13. Halaman eksekusi penyusunan daftar kata dan kemunculan hasil dalam peringkat 20 besar teratas.

Dari gambar tersebut, terlihat bahwa kata “Allah” adalah kata yang paling produktif atau paling banyak muncul di dalam Al-Qur'an. Sekali lagi, kata-kata tugas seperti *من* /min/, *في* /fi:/ dan beberapa kata tugas lain yang tampak dalam daftar tidak dihitung karena sifat dan kedudukannya yang suplementer dalam susunan kalimat ayat-ayat Al-Qur'an maupun dalam wacana tekstual yang umum. Yang menarik, misalnya, ada kata kerja yang paling produktif muncul dalam Al-Qur'an, yaitu *قال* /qa:la/ ‘berkata’ (mahluk) ‘berfirman’ (Allah) ‘bersabda’ (Rasul). Secara semantis, fenomena itu mengisyaratkan sesuatu terkait wacana dalam Al-Qur'an secara umum dan bisa didalami lebih lanjut berdasarkan konteks-konteks tertentu yang tersebar dalam berbagai ayat di dalam Al-Qur'an.

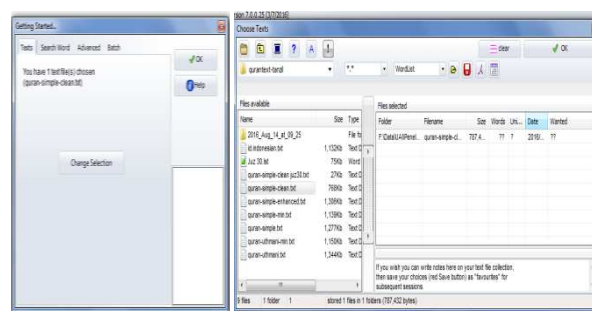
Setelah gambaran proses penyusunan daftar kata, berikut gambaran proses penyusunan konkordansi dalam Al-Qur'an.

- Buka aplikasi WordSmith, lalu klik menu Concord di bagian kiri atas halaman menu utama aplikasi, sampai muncul halaman awal Concord.



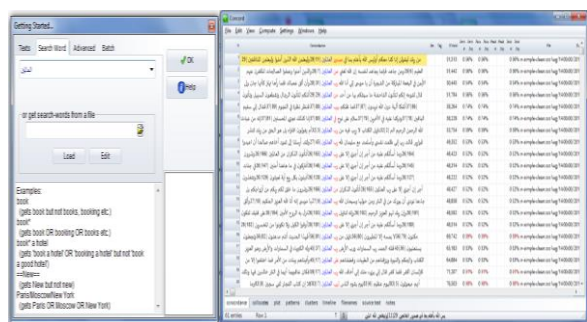
Gambar 14. Halaman awal aplikasi WordSmith dan menu Concord

- Buat file baru melalui menu *File* lalu submenu *New* maka akan muncul halaman berikut. Semula, menu *Change Selection* belum menunjukkan ada file yang dipilih. Setelah diklik, akan muncul halaman seperti di bagian kanan bawah berikut, lalu cari folder tempat file Al-Qur'an digital berada dan pindahkan ke halaman bagian kanan *Files selected*, lalu klik OK di bagian kanan atas halaman.



Gambar 15. Halaman menu dan pemilihan file yang akan dianalisis

- Setelah teks yang dipilih muncul di halaman semula, klik menu *Search Word*, maka akan muncul halaman seperti berikut.



Gambar 16. Halaman eksekusi penyusunan konkordansi dan kemunculan hasil dalam susunan 20 teratas.

Dari gambar tersebut, kita misalnya akan mencari konkordansi kata *العالمين* /al-'a:lamīn/ 'alam semesta'. Terlihat bahwa kata itu sangat dekat kata *رب* /rabb/ 'Penguasa' 'Tuhan'. Secara semantis, fenomena itu mengisyaratkan bahwa kata *العالمين* berkolokasi erat dengan kata *رب* /rabb/. Oleh karena itu, dapat disimpulkan bahwa kedua kata membentuk kolokasi berupa kata majemuk dan bermakna tunggal atau menyatu antara keduanya.

Demikian secara ringkas gambaran proses dan langkah-langkah penyusunan daftar kata dan konkordansi dalam Al-Qur'an dengan aplikasi WordSmith yang ternyata secara teknis mampu mengatasi masalah-masalah teknis ortografi nonbahasa Latin.

KESIMPULAN DAN SARAN

Berdasarkan gambaran pada bab-bab terdahulu, berikut dirumuskan simpulan dan saran dari penelitian ini.

1. Aplikasi yang paling representatif digunakan untuk membuat model file korpus Al-Qur'an digital adalah WordSmith. Itu karena aplikasi tersebut tidak hanya mampu mengatasi masalah teknis ortografi tulisan nonbahasa Latin, tetapi juga menyediakan fungsi-fungsi yang beragam dan sangat berguna bagi kepentingan analisis linguistik korpus.
2. Format *file* yang dapat diolah aplikasi tersebut haruslah dikonversi terlebih dahulu menjadi berekstensi *.txt, lebih khusus lagi, mode konversi *file* menggunakan *encoding* UTF-8.
3. Format isi teks digital berbahasa Arab yang akan diolah dengan aplikasi tersebut haruslah berisi karakter huruf sepenuhnya,

tanpa menyertakan tanda-tanda diakritik (*harakat*) yang dapat mengganggu teknis pengolahan *file*.

4. Proses membuat daftar kata dan konkordansi dalam Al-Qur'an dengan menggunakan aplikasi tersebut didasarkan sepenuhnya atas teknik-teknik yang tersedia dalam aplikasi secara memadai untuk mendapatkan hasil yang maksimal dan minim kendala teknis.
5. Secara khusus, model *file* digital semacam ini dapat digunakan oleh para peneliti pemula yang ingin menggunakan Al-Qur'an sebagai korpus data penelitiannya. Selain itu, mekanisme yang diuraikan dalam penelitian ini diharapkan dapat menjadi model bagi peneliti lain dalam hal tata laksana pembuatan korpus data berbahasa Arab.

Untuk lebih jauh lagi melanjutkan dan mengembangkan penelitian sederhana ini, berikut saran-saran yang diajukan.

1. Perlunya penelitian linguistik korpus secara bertahap namun berkelanjutan dan komprehensif terhadap Al-Qur'an khususnya untuk dihubungkan dengan bidang ilmu linguistik yang lain, seperti morfosintaksis, semantik, pragmatik, wacana, leksikografi, dan penerjemahan, terutama dalam kaitannya dengan penerjemahan Al-Qur'an ke dalam bahasa Indonesia.
2. Perlunya lebih digalakkan pendekatan linguistik korpus sebagai titik tolak dalam penelitian linguistik untuk memaksimalkan pemanfaatan data-data kebahasaan terkini yang sudah lebih banyak berbentuk digital.

DAFTAR PUSTAKA

- [1] Adolphs S, *Introducing Electronic Text Analysis - A Practical Guide for Language and Literary Studies*, Routledge, New York, p. 136-137, 2006
- [2] Cheng W, *Exploring Corpus Linguistics: Language in Action*, Routledge, Oxon, p.6, 29-30, 2012
- [3] [Wikipedia, List of Text Corpora](https://en.wikipedia.org/wiki/List_of_text_corpora), https://en.wikipedia.org/wiki/List_of_text_corpora, 2016, (Diakses pada 29 Juli 2016)
- [4] Sketch Engine, *Arabic TenTen Corpus*,

- <https://www.sketchengine.co.uk/artenten-corpus/>, 2015, (Diakses pada 29 Juli 2016)
- [5] Al-Sulaiti L, Arabic Corpora, http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm, 2010, (Diakses pada 29 Juli 2016)
- [6] Dukes K, The Quranic Arabic Corpus, <http://corpus.quran.com/>, 2011 [29 Juli 2016]
- [7] Adams A, QuranCode, <http://heliwave.com>, 2015 [29 Juli 2016]
- [8] Tanzil, Tanzil Quran Text Features, <http://tanzil.net/docs>, 2007 [29 Juli 2016]
- [9] Khan MK, Alginahi YM, The Holy Quran Digitization: Challenges and Concerns, Life Science Journal, p.156-164, <http://www.lifesciencesite.com>, 2013. (Diakses pada 29 Juli 2016).
- [10] Sharaf A, Atwell ES, QurAna: Corpus of the Quran annotated with Pronominal Anaphora, University of Leeds <http://www.textminingthequran.com/wiki/QurConcord>, 2012, (Diakses pada 29 Juli 2016)
- [11] Imad Z, Abdelhak L, New Quranic Corpus rich in Morphological Information, International Journal of Speech Technology, pp 339-346, 2016 (Diakses pada 29 Juli 2016)
- [12] Ismail D, Asyraf HA, *Mushaf Imla'i dan Implikasinya dalam Pembacaan Al-Quran (Imla'i Mushaf and It's Implication in Reading Al-Quran)*, 'Ulum Islamiyyah Journal, <http://ddms.usim.edu.my/bitstream/123456789/12106/1/ulum%20islamiyyah%2017%20-%203.pdf>, p. 43-55, 2016 (Diakses pada 29 Juli 2016)
- [13] <http://www.nooj4nlp.net/pages/nooj.html>
- [14] <http://neon.niederlandistik.fu-berlin.de/en/textstat/>
- [15] <http://www.monoconc.com/>
- [16] <http://www.laurenceanthony.net/software/antconc/>
- [17] <http://www.lexically.net/wordsmith/>
- [18] Alhaj A, New Theory of the Holy Qur'an Translation. A Textbook for Advanced University Students of Linguistics and Translation, Hamburg, Anchor Academic Publishing, 2015