Abstrak

Pembangunan korpus suara untuk pengembangan sistem Pengenalan Ucapan Kontinu Kosakata Besar (PUKKB) bahasa Indonesia berbasis *triphone* membutuhkan himpunan kalimat seimbang secara fonetik, yang mengakomodasi seoptimal mungkin *triphone* dan tanda baca. Namun, saat ini himpunan kalimat seimbang secara fonetik Bahasa Indonesia masih belum mengakomodasi *triphone* secara optimal karena diekstrak dari 500 ribu kalimat. Selain itu, himpunan kalimat tersebut sama sekali tidak memperhatikan tanda baca sehingga tidak bisa digunakan dalam pembangunan korpus suara yang mengakomodasi variasi intonasi. Fokus pada tugas akhir ini adalah untuk membangun himpunan kalimat seimbang berdasarkan fonem dan tanda baca. Himpunan kalimat seimbang ini dari hasil ekstrak 10 juta kalimat berbeda. Pengujian ini menunjukkan bahwa Semi LTM Greedy lebih fleksibel untuk menyeimbangkan jumlah *triphone* dan frekuensi kemunculannya dari pada algoritma sebelumnya *Modified LTM Greedy*.

Kata Kunci: *Triphone*, *Greedy*, Korpus Teks.