ABSTRAK

Sistem Pengenalan Ucapan Kontinu Kosakata Besar (PUKKB) Bahasa Indonesia berbasis silabel membutuhkan sebuah korpus suara yang seimbang secara silabel dan mengandung sebanyak mungkin silabel dan tanda baca. Saat ini, baru terdapat himpunan kalimat seimbang dengan jumlah silabel unik masih kurang karena baru menggunakan himpunan kalimat induk yang berjumlah sekitar 500 ribu kalimat. Penelitian ini bertujuan untuk mengatasi masalah tersebut dengan membangun sebuah korpus teks yang mempunyai jumlah silabel dan tanda baca yang maksimal dari himpunan kalimat induk berjumlah 10 juta kalimat. Dari hasil penelitian, didapatkan beberapa kemungkinan himpunan kalimat yang bisa digunakan sebagai data latih untuk sistem PUKKB.

Kata Kunci:*korpus teks*, *algoritma greedy*, *silabel*, *tanda baca*