

## ABSTRACT

Automatically generating a natural text that is perceived as grammatically correct remains a challenging task. The generated text must at least be coherent, accurate, and understandable. This research concerns the automatic generation of clitics in Pashto texts, since native Pashto speakers use clitics extensively in everyday conversation and writing. A clitic is a word or particle that cannot bear accent or stress, and phonetically leans on an accented adjacent word. Pashto language is spoken in Pakistan and Afghanistan. It is one of the several languages featuring clitics. There are two main types of clitics in Pashto: Second Position (2P) clitics and endoclitics. The linguistic behaviours of these clitics are studied and formalised into rules. The design of the Pashto clitic generation system is approached in two ways. In the first approach, system generates cliticised sentences from the semantic representation of the sentences. This system has been implemented using Combinatory Categorical Grammar (CCG). The second approach operates on the surface representation of sentences. It uses syntactic pattern matching rules for the identification and generation of clitics at sentence level. In this system, a text can be generated separately, so that after the text generation step, clitic generation rules can be applied to sentences as post-processing step. This system has been implemented in Python. The main advantage of this method is the separation of clitic generation task from the text generation task. The evaluation of the proposed solutions has been mainly constrained by the non-existence of morpho-syntactically annotated corpus, and language processing tools for Pashto. Notwithstanding, two independent corpora were developed. The first corpus contained semantic representations for generating 12 sentences based on Pashto CCG grammar. The second corpus consisted of 256 syntactically annotated sentences to evaluate the python-based clitic

generation system. The system is capable of generating all Pashto clitics including endoclitics, the most challenging clitic due to many constraints for its generation. All of the target sentences are successfully realised by the CCG grammar. The python-based Pashto clitic generator system achieves an accuracy of 89.62% on the test corpus. Incorrectly generated systems by the python-based generator have been fed to CCG generator to evaluate the agreement between the two systems. The accuracy achieved in this case is 87.5%.

**Keywords:** Clitic, pashto, clitic generation rules, combinatory categorial grammar

## ***Penjanaan Klitik Pashto***

### **ABSTRAK**

*Penjanaan teks tabii secara automatik dianggap betul dari segi tatabahasa terus menjadi tugas mencabar. Teks yang dijanakan mesti sekurang-kurangnya koheren, tepat, dan mudah difahami. Kajian ini menekankan penjanaan klitik secara automatik bagi teks dalam bahasa Pashto memandangkan penutur asli bahasa Pashto menggunakan klitik dengan meluas dalam perbualan seharian dan penulisan mereka. Klitik ialah perkataan atau partikel yang tidak menekankan telur bahasa dan secara fonetik bersandar pada perkataan aksen bersebelahan. Secara amnya, bahasa Pashto dituturkan di Pakistan dan Afghanistan. Bahasa Pashto merupakan salah satu bahasa yang mempunyai ciri-ciri klitik. Terdapat dua jenis klitik utama dalam bahasa Pashto iaitu klitik posisi kedua (2P) dan endoklitik. Tingkah laku secara linguistik terhadap klitik ini dikaji dan diformalkan ke dalam peraturan. Terdapat dua jenis pendekatan digunakan dalam reka bentuk sistem penjanaan klitik bahasa Pashto ini. Pendekatan pertama adalah sistem menjanakan ayat yang berklitik dari ayat perwakilan semantik. Sistem ini dilaksanakan menggunakan pendekatan "Penggabungan Tatabahasa Berkategori". Pendekatan kedua pula mengendalikan perwakilan pada permukaan ayat. Ia menggunakan peraturan pepadanan corak sintaks untuk pengenalpastian dan penjanaan klitik pada peringkat ayat. Dalam sistem ini, teks boleh dijanakan secara berasingan supaya selepas langkah penjanaan teks, peraturan penjanaan klitik boleh diaplikasikan pada ayat sebagai langkah pasca pemprosesan. Sistem ini telah dibangunkan menggunakan Python. Kelebihan utama kaedah ini ialah pemisahan bagi tugas penjanaan klitik dari tugas penjanaan teks. Penilaian ke atas cadangan penyelesaian telah dikekang oleh ketidakhadiran korpus beranotasi morfologi-sintaksis dan perkakas*

*pemrosesan bahasa untuk bahasa Pashto. Walau bagaimanapun, dua korpus tak bersandar telah dibangunkan bagi tujuan penilaian tersebut. Korpus pertama mengandungi perwakilan semantik untuk menjanakan 12 ayat berdasarkan tatabahasa "Penggabungan Tatabahasa Berkategori" Pashto. Korpus kedua pula terdiri daripada 256 ayat beranotasi dari segi sintaksis bertujuan untuk menilai sistem penjanaan klitik berasaskan Python. Sistem ini berupaya untuk menjanakan kesemua klitik bahasa Pashto termasuk endoklitik, iaitu klitik yang paling mencabar untuk dijana berikutan pelbagai kekangan dalam penjanaannya. Kesemua ayat sasaran berjaya direalisasikan oleh tatabahasa "Penggabungan Tatabahasa Berkategori" Pashto. Sistem penjana klitik bahasa Pashto berasaskan Python telah mencapai ketepatan 89.62% pada korpus yang diuji. Dari hasil janaan tersebut, kesemua ayat yang salah dibekalkan kepada penjana "Penggabungan Tatabahasa Berkategori" untuk menilai kesesuaian antara kedua penjana. Ketepatan yang dicapai dalam kes ini adalah 87.5%.*

**Kata kunci:** *Klitik, pashto, penjanaan peraturan, penggabungan tatabahasa berkategori*