Exploring AI-Based Syntax Parsing for Understanding Iraqi Arabic-English Mixed Sentences

Prof .Dr. Kareem Lazim
Prof.Dr. Fatima Raheem
University of Misan

Abstract

The rapidly accelerating effect of code-switching in a linguistically diverse community, such as an urban place in Iraq, calls for more advanced tools for analyzing mixed-language constructs. This research evaluates the potential of artificial intelligence-based syntax parsing for understanding mixed Iraqi Arabic-English sentences. Using a corpus of actual code-switched sentences taken from urban Iraqi speakers, the research identifies how AI algorithms, specifically neural network-based natural language processing (NLP) models, could parse syntax accurately. It assesses the performance of AI parsers across complex linguistic phenomena such as morphosyntactic alignment, word order alternations, and context-dependent lexical choices that characterize code-switching between Iraqi Arabic and English. A comparative study on performance regarding accuracy, recall, and adaptability of AI-based parsers trained on different datasets: monolingual and mixed-language is performed to achieve this. This study also investigates the sociolinguistic aspect of code-switching, looking at how cultural norms and linguistic choices influence syntactic patterns. As per these reports, the AI-based syntactic parsing will demonstrate an increasing tendency towards much better decoding accuracy relative to structural complexity of the bilingual sentences, as trained on mixed-language data. Such results leave a need for the development of application-specific AI tools for linguistic analysis, language learning, and multilingual communication technologies in these linguistically rich environments.

Keywords : AI, syntax, NLP, code-switching

استكشاف تحليل البنية النحوية باستخدام الذكاء الاصطناعي لفهم الجمل المختلطة بين العراقية والإنجليزية

البروفيسور د. كريم لازم
البروفيسور د. فاطمة رحيم
جامعة ميسان

الملخص

الأثر المتسارع لظاهرة التحول اللغوي (code-switching) في مجتمع لغوي متنوع، مثل المدن العراقية، يتطلب أدوات أكثر تقدمًا لتحليل البنى اللغوية المختلطة. يقيّم هذا البحث إمكانيات تحليل البنية النحوية باستخدام الذكاء الاصطناعي لفهم الجمل المختلطة بين اللهجة العراقية واللغة الإنجليزية. باستخدام مجموعة

بيانات مأخوذة من جمل فعلية تم تحويلها لغويًا من متحدثين عراقيين في المناطق الحضرية، يحدد البحث كيفية قدرة خوارزميات الذكاء الاصطناعي، وخاصة النماذج المعتمدة على الشبكات العصبية في معالجة اللغات الطبيعية (NLP)، على تحليل البنية النحوية بدقة. يقيم البحث أداء محللات الذكاء الاصطناعي في التعامل مع ظواهر لغوية معقدة مثل التوافق المورفولوجي-النحوي، التغيرات في ترتيب الكلمات، والاختيارات المعجمية المعتمدة على السياق، التي تميز التحول اللغوي بين العراقية والإنجليزية. لإجراء ذلك، يتم إجراء دراسة مقارنة لأداء المحللات النحوية المعتمدة على الذكاء الاصطناعي من حيث الدقة، والاسترجاع، والقدرة على التكيف، وذلك بناءً على بيانات تدريبية أحادية اللغة ومختلطة اللغة.كما يبحث هذا البحث الجانب السوسيولغوي للتحول اللغوي، من خلال النظر في كيفية تأثير الأعراف الثقافية والاختيارات اللغوية على الأنماط النحوية. وفقًا لهذه التقارير، من المتوقع أن تظهر معالجة البنية النحوية باستخدام الذكاء الاصطناعي ميلًا متزايدًا نحو دقة أفضل في فك تشفير الجمل الثنائية اللغة بما يتناسب مع تعقيدها الهيكلي، عند التدريب على بيانات مختلطة اللغة. تؤكد هذه النتائج الحاجة إلى تطوير أدوات ذكاء اصطناعي مخصصة للتطبيقات اللغوية، تعليم اللغات، وتقنيات التواصل متعدد اللغات في هذه البيئات الغنية لغويًا.

الكلمات المفتاحية: الذكاء الاصطناعي، بناء الجملة، البرمجة اللغوية العصبية، تبديل التعليمات البرمجية

1.1 Introduction

One of the most distinct characteristics of multilingual societies is code-switching-involving a speaker's fluent transition between two or more languages in a single utterance, sentence, or phrase. This linguistic behaviour particularly manifests in Iraqi bilinguals - those who are able to speak Arabic and English-albeit more among urban centers where cultural hybridity, globalization, as well as exposure to education in English, strongly influence it. Bilingual communities create more elaborate codeswitching which does not only reflect individual linguistic competencies, but also derive from broader social norms, cultural expectations, and pragmatic functions of language in daily communicative interaction. This practice defies then traditional linguistic analysis, as it has to do with syntactic, morphological, and semantic mixing that is not always consistent with the grammatical prescriptions of either language. 'Mentioned language changes are not likely limited to the alternation or interpolation of a word, but often comprise, all at once, switchings occurring directly within the entire utterance, which, then, becomes an utterances- that seem mixed touches of both languages reflecting both innovations in languages and cultures. Bilingual practices of this kind require methods of analysis that greatly transcend the monolingual paradigm of understanding.Recent It is the most recent development of artificial intelligence (AI) and natural language processing (NLP), which could pave the way towards decoding the intricacies of code-switching. Syntax parsing entails deconstruction and organization of grammatical structures such that one will uncover the rule that underlies their sentence construction. Current AI-based parsers are effective in analyzing monolingual texts; however, much remains to be done when one comes to parsing code-switched inputs. It is because these parsers silently fail with the fluid nature of bilingual language use, such as shifting syntactic structures, borrowing, and morphological adaptations.

Thus developing a specific AI tool to parse and analyze the code-switched sentences fills this gap. Such tools would be for Iraqi Arabic-English bilinguals taking into account the diversified attributes of both Iraqi Arabic and English linguistic features as well as those sociolinguistic factors that underpin their combination. It suffices to say that one would need datasets which are culturally and contextually relevant: another essential part of their development to add up the accuracy and reliability of the parsers concerned. More so, creating models that could recognize the grammatical and functional roles of code-switching would open a new window into bilingual language use, especially cognition, identity, and communication effects.. Ultimately, advancing AI-based syntax parsing for code-switching will not only fill a critical gap in linguistic research but also offer practical applications in education, language learning, and technology development. For Iraqi Arabic-English bilinguals, such innovations hold the potential to foster better understanding, support bilingual education, and enhance tools for natural language interfaces that reflect the reality of multilingual communication.

1.2 Problem Statement

The lack of instruments that can adequately analyze code-switched syntax poses significant barriers for linguists, educators, and developers of technological applications for language. Current syntactic parsers mainly trained on monolingual data and try to analyze the unique grammatical patterns of Iraqi Arabic-English mixed sentences, which build up beyond rest-cased features, such as basic English terminology fused into Arab syntactic frames or shifts in word order. Current tools cannot identify above all. This limits linguistic investigation, in which it is important to know the syntactic and pragmatic dimensions of code-switching such as provided by existing educational strategies concerning the realities of bi-cultural communication in the classroom and the development of bilingual realities manage educational strategies within classrooms. Besides, there is still no development of multilingual technologies such as virtual assistants and language-learning platforms for the bilingual audience because their system cannot parse mixed-language inputs. This study will help shed light on these gaps by adapting AI-based syntax parsers to deal with greater accuracy in Iraqi Arabic-English code-switching. Considering the uncertainty of switching between two languages by an individual, it is a reasonable proposition that there should be an awareness in the embedded languages or heads in the switching phrases. This notion is furthered by the development of context-aware parsers through annotated datasets that mimic linguistic differences of code-switching speech and might apply advanced machine learning techniques, from deep learning to transfer learning. These tools will identify syntactic patterns combined with social and functional clasps for code-switching. The results of this study could also serve to advance the theoretical understanding of bilingualism while also providing application use cases in teacher training and in more inclusive multilingual technologies tailored to the needs of Iraqi Arabic-English speakers.

1.3 Objectives of the Study

The primary objectives of this study are as follows:

1. To investigate the syntactic characteristics of Iraqi Arabic-English mixed sentences.
2. To evaluate the performance of existing AI-based syntax parsers on code-switched data.
3. To develop and test a framework for adapting AI parsers to handle mixed-language syntactic structures.
4. To examine the implications of AI-based syntax parsing for linguistic research and multilingual applications.

1.4Research Questions

This study seeks to answer the following questions:

1. What are the key syntactic features of Iraqi Arabic-English mixed sentences?
2. How do existing AI parsers perform when analyzing code-switched data?
3. What modifications are necessary to enhance the accuracy of AI parsers for mixed-language sentences?
4. What are the potential applications of AI-based syntax parsing in linguistics and multilingual technologies?

1.5 Significance of the Study

More and more applied linguistic and computational linguistic associated issues are being addressed by research in code-switching analysis. This will assure that researchers studying bilingualism and educators interested in language pedagogy would find valuable insights obtained. Furthermore, the finding can also be useful for the development of AI-driven tools that will allow implementation of multilingual communication technologies such as chatbots, virtual assistants, and automated translation systems.

1.6 Scope and Limitations

The study focuses on Iraqi Arabic-English bilinguals in urban areas, namely in Maysan emphasizing syntactic analysis rather than phonological or semantic aspects of code-switching. While the research aims to adapt AI parsers for mixed-language sentences, it does not involve creating entirely new parsing algorithms. The study's findings may have limited generalizability to other language pairs but offer a framework for future research in similar contexts.

2: Literature Review

2.1 Code-Switching: Definitions and Theoretical Perspectives

Code-switching is considered to be a very complex phenomenon, according to which the conversation involves either one sentence or else two different languages. It is considered complex as it is usually influenced by some factors, such as the sociolinguistic, cognitive, and structural, but the influences of these factors usually depend on the context in which the speaker is found, as well as the language proficiency of the speaker, along with the communicative goals in hand (Poplack, 1980). Switching from one language to another is not a random act. It is rather influenced by several motivators, including the need for clarity, reinforcement of social identity, and expression of culture. Code-switching has been described by Al-Khatib and Sabbah (2008) as a phenomenon which can be employed to nuanced meanings, group membership, and social relationship negotiation between bilingual communities where speakers of Iraqi Arabic and speakers of English thrive together. This is true in contexts where there are bilingual people wishing to either mark certain aspects of their identity or modulate their speech according to formality or informality of the situation. Two theoretical frameworks are particularly useful in understanding code-switching:

1. Sociolinguistic Approach : guests have to do with the ways in which different social factors, including the contexts or audience and even cultural norms, play their parts in shaping language choice. In fact, the bilingual hispanics make their decisions to switch language becouse of various other social factors including the speaker role, the language proficiency perceived from the interlocutor, and the general societal attitude towards the languages involved. In the case of Iraqi Arabic-English bilinguals, switching between languages may reflect sociocultural meanings ranging from solidarity with specific social groups to an index of proficiency in both languages with different communicative aims.

2. Grammatical Approach- This approach examines the difference between languages with respect to the grammar and looks at the various rules under which languages will overlap as far as grammar is concerned. It mostly focuses on

morphosyntactic rules regarding language alternation like word order, verb-noun alignment, and subject-verb agreement. Myers-Scotton (1993) states that the codeswitching may be subject to certain syntactic rules that will facilitate changing in the integration of different language components. This is also a relatively fruitful approach in studying the way that lexical items of Iraqi Arabic and English are incorporated in sentences or phrases formed by bilingual speakers while keeping overall coherence and fluency despite different grammatical structures of the two languages. Studies concerning the phenomenon of code-switching when it comes to Iraqi Arabic and English have recognized certain syntactic patterns that could be peculiar to this bilingual interaction, which would include a special verb-noun alignment, specific instances of subject-verb agreement, and so on that are genuinely characteristic of the bilingual context as they are indicative of the speakers' ability to navigate with ease the structural nuances of both languages at the same time (Holes, 2004). But still, it created a significant gap in understanding what it means by the syntactic patterns in being applied or modeled within a computational context while carrying substantial theoretical and practical weight in what those findings discovered. There is an emphasis on how such patterns would require research into the very computational modeling of such code-switching because they may provide insights into artificial intelligence processing of natural language, machine translation, or, even further, any other area needing efficient interpretation and analysis of bilingual language data. Yet, as bilingualism goes on increasing worldwide, especially in countries like Iraq, where English and Arabic are both used, it becomes important not only in linguistic theory but also in applied computational linguistics-the mechanics of code-switching.

2.3 Syntax Parsing in Linguistics

Parsing syntax fundamentally entails a productive study of the grammatical composition of a sentence for analyzing the inter-relationship between different components such as phrases, clauses, and their syntactic dependencies. Generating a syntactic tree or structure that would sufficiently describe the underlining grammatical organization of a given sentence is what is targeted by this method. Most of the traditional syntax parsers used rule-based models or statistical models for processing natural language. Rule-based parsers depend upon predetermined grammatic rules to be advanced, which builds the structural formation of the sentences. A Probabilistic Statistical Parser is a statistical parser that predicts the syntactic structures or configurations based on their training using some enormous datasets of probabilities. Several of this method proved to work effectively for many types of text but fail where code-switched entries are in effect with two or more languages within a single sentence or discourse. Code-switching does bring about a very challenging environment for traditional parsing involving unpredictable syntactic changes that happen while switching from one language to another by the speaker. Factors inducing this change include an entirely different linguistic structure, social context of this alteration, and the necessity to balance two grammars distinct from each other. As indicated by Jurafsky and Martin (2023), code-switched data is often too irregular and variable for rule-based or statistical models to handle effectively, denoting that these special features had better resort to employing special parsing models dedicated to the unique context of code-switched sentences.

.Several key challenges arise when attempting to parse mixed-language sentences, particularly when dealing with language pairs like Iraqi Arabic and English, which exhibit strikingly different grammatical properties. These challenges include:

1. Morphological Complexity: Iraqi Arabic is a highly morphologically rich language with complex inflectional morphology. This means that the same word in Arabic may change its form by grammatical features such as tense, case, gende and number, making it a curious inflected language. For example, Arabic verbs are highly inflected in both tense and aspect, with nominal constructions whose argument forms are marked for case and gender. English, on the other hand, tends to be simpler with regard to morphological structures such that fewer inflections apply and grammatical relationships are indicated rather by auxiliary verbs accompanied by word order. But when such a mixture occurs, they become very much problematic for traditional parsers to distinguish between such complex morphological forms of Iraqi Arabic and simpler forms of English, especially when inflections and syntactic markers overlap or would even be transferred incorrectly across languages.

2. Word Order Variations: One of the most significant syntactic differences between Iraqi Arabic and English is regarding the order of words. It is mainly seen that Iraqi Arabic uses the verb-subject-object ordering, while English has more of a subject-verb-object. This brings those code-switching utterances into diluted parsing ambiguities with other types of code switches as well. But some sentences would require double parsing: one, say a sentence has a double life with both Arabic and English items, for it can become ambiguous, whether it will have a VSO parsing or SVO. Such variations are severe to parsers trained on traditional monolingual data as they can misinterpret or misparse anything.

3. Lexical Ambiguity: Lexical ambiguity is where a term can fall into either of the two languages within a code-switched sentence, given a particular context of being used. Code-switching between Iraqi Arabic and English utilizes many words that may contain meanings from both languages and change their syntactic behavior according to the language context. For example, an English noun that is inserted into an Arabic sentence will carry different grammatical properties than it would have carried in its purely English construction. Besides, there are also words in similar forms in the two languages but are completely different in meaning, which complicates parsing further. For a traditional parser, it is an effortful task concerning determining the language a word belongs to and how it would be used within an appropriate syntactic structure of the sentence.

2.4 AI and NLP in Syntax Parsing

The advances and applications of machine learning (ML) and natural language processing (NLP) techniques have been at the core of revolutionizing the way linguistic analyses are performed through artificial intelligence. Such advances have resonated with and shifted the procedures of automating and improving many linguistic tasks, including syntactic parsing, semantic-level understanding, sentiment analysis, and, indeed, language generation itself. Artificial work processing based on AI using neural network models is one of the most significant innovations in language processing today that has transformed the efficiency and correctness of parsing one language. One such area of work has been the introduction of transformer models-believed to have set new standards in handling one language tasks in the linguistic world. Transformers are an architecture proposed by Vaswani et al. in 2017. They process input data in a parallel manner using self-attention mechanisms and therefore allow the model to capture longer dependencies and contextual relations in texts more efficiently. In that regard, they have succeeded massively, especially concerning given monolingual-niche NLP tasks such as syntactic parsing, named entity recognition, and machine translation. The architecture in question allows the semantical contextualization of single words/phrases in a sentence, which is a very important step in understanding the meaning of language in context. This is mostly why transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), are suited for syntax- and semantics-nuanced understanding tasks. The adoption of transformer-based models has proven rather challenging when it comes to code-switched syntax, and one of such notions is code-switching that allows two or more languages to mix within a conversation or a sentence. It adds extra layers of complexities within themselves, making them not perform as efficiently as expected. The traditional transformers are BERT as a model for monolingual data, and these models are just ineffective in capturing the unique syntactic and semantic features present in code-switched sentences. This is due to the fact that it results in unpredictable variations in syntax, morphology, and lexical choice between languages while switching a language and hence makes the standard models incapable of producing accurate and coherent mixed-language data outputs. Emerging from this gap, recent research targets modifying and improving transformer models for multilingual and code-switched scenarios. This can be elaborated as: two key developments in the lines of Multilingual BERT (another development apart from the BERT-based code-switching models) that serve to enrich the framework for mixed-language data representation.

BERT and multilingual BERT (Devlin et al., 2019): Being a transformer model, BERT is trained on large corpora to create bidirectional contextual embeddings, meaning obtaining vector representations that capture the meaning of a word together with that of the surrounding context. This model, however, is called multilingual BERT because it was trained only on strings from several languages at once, thus enabling it to comprehend and produce output in any of the languages. mBERT has also proven its worth in tasks like cross-lingual language understanding and machine translation, because it basically shares the same benefits having good multilingual data for their performance. One of the strengths with mBERT is applying the notion of making relationships for all the languages in common space in a single model, perhaps for processing inputs of multiple languages, but not always work in the best possible ways while code-switching them.

2.5 Iraqi Arabic-English Code-Switching: Linguistic Features

The existing diversification of factors shaping English-Iraqi Arabic bilingualism has historical, social, and educational underpinnings that have gradually made English a language increasingly present in relatively many contexts. English has over time formed part of key areas, comprising education, media, technology, and business, and has actively fostered the access of both languages into everyday use. Bilingualism is not only a phenomenon of contact between languages but also a reflection of sweeping changes in society-crucially linked with the ways in which the Iraqi society operates with the global trend, international institutions, and modernization. This is hence why they extensively code-switch between the respective Iraqi Arabic and English to interact more effectively in their urban and educated settings. Some of the metafunctions that characterize code-switching of Iraqi Arabic-English will be as follows:

1. Lexical Borrowing: Undoubtedly, one of the most outstanding characters of bilingualism in Iraqi Arabic and English is the absorption of English terms, particularly technical, scientific, and technological vocabulary, into Arabic sentences. This lexical borrowing occurs in certain domains, most notably those areas in which English has a specialized term that does not yet find an equivalent in Arabic or in which the English term is felt to be more modern, prestigious, or

efficient. It is common to hear the English form, e.g., "internet," "computer," or "marketing," used within an Arabic sentence without translation or with little adjustment to Arabic phonology (Jassem, 2012), as with those related to computing, business, or science. Such borrowed terms constitute lexical gaps in the Arabic lexicon and testify to the increasing influence of English in the globalized world. However, the integration of these terms is very fluid because the speaker would switch languages or maybe adopt the kind of term to fit in with the context of communication and the specificity of the particular lexicon needed.

2. Structural Integration – Besides lexical borrowing, another phenomenon of structural integration is the integration of an English syntactic structure into Arabic sentences or vice versa. There are many phenomena, like, to make Arabic sentences follow the English word order pattern (subject-verb-object or SVO), or to embed English syntactic elements such as auxiliary verbs or prepositions in the Arabic sentence. In contrast, sentences in English could at times carry Arabic influences in their structure. This could be due to the competence of the speaker in Arabic being much higher, which may sometimes influence English usage. All this serves to show that bilingual speakers know how to manipulate syntactic rules of both languages and can create such hybrid sentences taking the best out of both worlds. It is known that such integration of syntax happens unconsciously, where the speaker just turns from one language system to the other, depending on the communicative context. (Albirini, 2016). It also highlights the flexibility of bilingual speakers, who are able to adapt their language use to the demands of their social, cultural, and communicative environments.

3. Switch Points: For Iraqi Arabic-English bilingualism, code-switching occurs typically at precise points in speech clauses or between subjects and predicates in the sentence. This is consistent with Poplack's (1980) Equivalence Constraint, which postulates that shifting in language is more likely to take place at areas whereby both tongues afford syntactically parallel forms. For instance, switches between clauses in each clause have independent syntactic units rendering a smooth shift from one language to another. So too between subject and predicate, as both the languages afford comparable grammatical structure for smooth transition. These points of switching become representative of the speakers' rhythms of doing such kind of scattered speech, where one language is switched for purposes of communicating a particular idea or clarifying a point.

3: Methodology

3.1 Research Design

This research uses mixed-methods: both qualitative and quantitative to comprehensively investigate the code-switching between Iraqi Arabic and English, and its syntactic attributes. This combination is necessary for adequately capturing the nuanced linguistic features of mixed-language sentences. It must also include the computational challenges that this complex syntax poses in bilingual discourse. It includes performing a linguistic analysis to identify and account for the major syntactic patterns present in code-switched sentences as well as evaluating how well AI-based syntax parsers process such data. This dual approach is essential because through such, one is able to derive a comprehensive understanding of the linguistic phenomena involved along with the technological challenges that need to be overcome in parsing mixed-language data.

3.1.1 Qualitative Component

The main qualitative part of the research attempts to provide an exhaustive linguistic analysis of the Arabic-English code-switching sentences in the Iraqi corpus. It attempts at determining the syntactic structures that typify mixed-language discourse. It investigates such familiar patterns of sentence structure as word order, subject-verb agreement, and verb-noun alignment, most often disrupted or changed when switching from one language to the other. The two languages, Iraqi Arabic and English, have different syntactic frameworks; Iraqi Arabic is typically VSO (verb-subject-object), conversely, English is SVO (subject-verb-object). Such comparative insights will be gotten in understanding how the two systems blend in a code-switched sentence with respect to syntactic adjustments made by the bilingual speaker.

Besides identifying this structure, analysis takes an additional step, exploring the language mechanisms that underlie code-switched speech. It also studies points of switching within a sentence by bilingual speakers-whether they would switch at the clause boundary, switch between phrases, or even alternative switching within a phrase. Such work is imperative toward understanding the syntax of bilingual contexts, particularly language dominance, the complexity of the lexical items involved, and the syntactic boundary between languages.

3.1.2 Quantitative Component

The quantitative component is the assessment of the performance of AI-based syntax parsers in the generalization property of their handling mixed-language sentences, just as any other sentence is monolingual. This component is necessary for testing the adaptability of these AI models for the syntactic complexity found in code-switched data. AI parser performance is evaluated on how well they can parse code-switched sentences and assign accurate syntactic labels to the components of those sentences. This evaluation uses standard machine learning metrics such as accuracy, precision, recall, and F1-Score. These retain the nature of "human" syntactic structure but will randomly ask for comparisons against the human annotated syntactic structures of the corpus for the performance of the model. The accuracy checks whether the parsed sentence is correct or not; while precision and recall will tell how well it picks out joins related to specific syntactic relationships, such as subject-verb agreement or word-order. The F-measure, which balances precision and recall, gives a composite indication of the parser's performance with respect to mixed-language syntax. Based on this comparison, one may tell if the changes made to adapt the parser for such cases have worked well or not. This kind of comparison points out the difficulties that face by the AI-based systems where they run into while trying to understand non-standard, mixed-language input when that is compared to the structured, monolingual sentences.

3.1.3 Integration of Qualitative and Quantitative Approaches

Summing up both qualitative and quantitative methods leads to a whole picture of the Iraqi Arabic-English code-switching dynamics and data analyses from AI syntactic parsing features. On the other hand, the quantitative one looks into how such tools confine the applicability and accuracy of syntactic parsing in practice-aspects. Such a combination ensures that the study identifies the major linguistic features of a mixed-language sentence as well as the treatment of AI-based ones in regard to processing such features. Such an entry point will eventually lead to a triangulation of the results generated from both components, thus producing an overall understanding of the syntactic properties of Iraqi Arabic-English code-switching. This integrated approach also informs the improvement of AI-based syntactic parsers, thus aiding in the development and building of better tools for multilingual language processing. Furthermore, the findings can be applied to theoretical linguistics in exposing the complexity of bilingual syntax and the adaptability of language systems to the codeswitching condition.This mixed-methods design is particularly valuable for addressing the complex and multifaceted nature of code-switching, where both linguistic expertise and computational analysis are necessary to provide a comprehensive understanding of the phenomenon. By combining these two methodologies, the study aims to bridge the gap between theoretical linguistics and applied computational linguistics, offering valuable insights into the syntactic mechanisms of bilingual discourse and the potential of AI technologies to analyze such discourse effectively.

3.2 Data Collection

3.2.1 Corpus Selection

There is a meticulously compiled corpus of Iraqi Arabic-English code-switched sentences, which serves as primary data for this study, coming from a wide range of sources so as to reflect the natural and dynamic use of language in bilingual settings. The corpus is thus intended to contain all the many ways in which Iraqi Arabic and English can come together in everyday interactions- be they written or spoken. Since the socio-linguistic context of urban Iraqi communities is highly characterized by bilingualism, the corpus brings to attention a vast number of various ways that code-switching occurs, as well as between very different contexts and registers.

It well captures the variety and complexity of code-switching in between Iraqi Arabic and English, and it has been thought intensively about different forms of code-switching. The data will consist of both intra-sentential (within-sentence) and inter-sentential (between-sentences) switches, so that - the systems of the two languages can be well viewed with regard to how they coexist and interact in the same communicative event.

3.2.2 Conversational Data

One key source for the corpus is conversational data collected from informal dialogues between bilingual speakers of Iraqi Arabic and English. These dialogues were transcribed from recordings of conversations that took place in everyday settings such as family gatherings, social events, and casual interactions in urban areas. Conversational data are particularly valuable for this study because they reflect the natural use of code-switching in spontaneous communication. In these dialogues, bilingual speakers often shift between languages based on factors such as topic, familiarity with the interlocutor, or the need to convey specific cultural references. The spontaneous nature of this data offers rich insights into the fluidity and adaptability of syntax in code-switched speech, providing examples of intra-sentential switches where a single sentence or clause may seamlessly incorporate both languages. The conversational data not only provides instances of code-switching but also captures the socio-pragmatic contexts in which code-switching occurs. This aspect is particularly important because it highlights the social functions of code-switching, such as marking identity, signaling familiarity, or navigating social hierarchies within bilingual communities.

3.2.3 Social Media Data

Another substantial source for the corpus is social media data: publically available postings, comments, and discussions found in Facebook, Twitter, and Instagram. It is the medium through which increasingly volatile communication is manifested among bilingual communities; indeed, it serves as a mine of written data which reflects the integration of Iraqi Arabic and English. Posts and comments on social media sites are ever dynamic combinations of both languages, as speakers mix popular colloquial expressions, vernacular slang, and technical words from English into their predominantly Arabic discourse. Social media will offer a one-of-a-kind opportunity to probe how bilingual speakers code-switch between languages in writing, which may be different in syntax and switching points from what is thought of as spoken discourse. In social media, the whole setting provides the possibility that code-switching may be very situational-based-to a humor value or an indication of joining-in on global or Western taste-oriented discourse. This source would thus contribute to the research as a means of discerning how bilingual speakers manage to navigate the blurred lines between formal and informal registers in writing and how such navigations affect sentence formation and syntactic alignments in either language.

3.2.4 Academic Contexts

The third important data source is the academic context as it elaborates on educational materials and textbooks used within Iraqi schools, where English is taught into the integrated Arabic discourses. English is one of the leading foreign languages in the Iraqi education system, with heavy borrowings in the academic texts, lectures, instructional materials, and especially in subjects like science, technology, and business. These educational materials mostly contain code-switching and the terms or concepts in English that are incorporated into an otherwise Arabic language sentence. This would be considered inter-sentential code-switching, where a sentence in Arabic could actually warrant an English term or phrase or vice versa. Instances of code-switching would usually happen when the speaker or writer puts forward a specific academic term that does not find equivalent in Arabic or that addresses a more globular or specialized audience.

It contains not only textbooks, but also written materials from online academic consensus, where pure formal-informal language more often than not manifests as the new trend of including English in the academic discourse of Iraq. Academic contexts are important sources of evidence for examining the ways in which bilingual speakers experience the realities of coexisting languages within those same categories of more formal and structured environments like lectures, research papers, and discussions in educational venues.

3.2.5 Corpus Composition

The final set of compiled data for this research study was more-or-less 2,000 sentences with an even distribution for the code-switching forms-within-sentence and between-sentence switches. The integration of both forms is needed because actual bilingual discourse exhibits a full range of syntactic structures. Intra-sentential switches (whereby speakers switch languages within a sentence) are valuable for examining the grammar rules from both languages as they merge into one structure. They reveal, in most cases, how they adjust the sentences according to the differences in the Iraqi Arabic and English syntactic structures, especially concerning the verbal alignment with the noun, subject-verb agreement, and word order.

3.3.Linguistic Features

The collection of data is mainly aimed at capturing syntactic features of code-switched sentences, namely:

Word Order: Variability between word orders of Arabic for VSO versus English for SVO.

Morphosyntactic Alignment: Concords among subjects, verbs, and objects in hybrid sentences.

Switching Positions: Points of switching locations, particularly into clauses and phrases.

And this data serves as the foundation for the linguistic analysis as well as the AI-based parsing evaluation.

3.4 AI-Based Syntax Parsing Tools

3.4.1 Selection of Parsing Models

For the AI-based syntax parsing, two primary models were chosen:

Universal Dependency Parser: A neural network-based model trained on multilingual data, designed to identify syntactic relationships in sentences across languages (Andersen et al., 2020).

Multilingual BERT (mBERT): A pre-trained transformer model capable of handling multiple languages simultaneously. mBERT was specifically chosen for its effectiveness in processing code-switched sentences, as demonstrated in recent studies on mixed-language data (Pires et al., 2019).

Both models are adapted to code-switched data through fine-tuning using the collected corpus of Iraqi Arabic-English sentences.

3.4.2 Preprocessing and Data Augmentation

Prior to training the AI models, the code-switched sentences undergo preprocessing, including:

Tokenization: Breaking down sentences into individual tokens (words, punctuation, etc.).

Language Tagging: Labeling tokens with their respective language (Arabic or English) to help the parser differentiate between syntactic rules.

Data Augmentation: To enhance model robustness, additional artificially generated code-switched sentences are incorporated, expanding the diversity of linguistic structures.

3.5 Research Procedures

3.5.1 Linguistic Analysis

The linguistic study aims to identify the commonly occurring syntactical structures that are code-mixed in the corpus, such as word order, syntactic agreement, and switch points. The following stages are followed:

1. Annotation: The annotation is done by linguists specialized in Iraqi Arabic and English who point out switching codes and syntactic features like subject-verb agreement and phrase structures.

2. Identification of Patterns: Referring syntactic patterns are then qualitatively analyzed to identify general structures and deviations from monolingual syntactic norms.

3.5.2 AI Parser Training and Evaluation

AI models are trained and evaluated using the following steps:

1. Training: The corpus coded in mixed language is preprocessed and divided into training (80%) and validation (20%) datasets for further enhancement of acquisitions according to mixed-language syntax with these datasets.

2. Evaluation: After finishing the training, the models are evaluated using a different test dataset that has never been seen by any of the code-switching sentences. Their performance is also evaluated based on standard metrics: accuracy, precision, recall, and F1-score.

3. Comparison with Monolingual Data: The AI models designed for code-switched data were made to run and compared with their counterparts designed for monolingual Iraqi Arabic and English data to know how efficient these adaptations were.

3.6 Data Analysis

3.6.1 Qualitative Analysis of Syntax

It involves examining common syntactic characteristics of the code-switched sentences. The identifiers are as follows:

Switch Points: The occurrences and form in which switching happens concerning the clause boundary, phrase structure, and syntactic dependencies.

Structural Features: Here, the focus is on the word order, agreement, and other syntactic patterns with the view of identifying language-specific tendencies in the mixed sentences.

1. Switch Points

Unlike other languages, switch points could be internal points of a sentence or phrase, as well as points in discourse at which a change occurs from one to the other. Such an analysis would reveal the points of code-switching favourite within intra-sentential and inter-clause switching, or even in a phrase within linear clause switching.

Findings:

Intra-sentential switches are the most frequent form of code-switching observed, accounting for 68% of all switches in the corpus. These occur primarily at the syntactic boundaries of noun phrases, verb phrases, and prepositional phrases.

Inter-sentential switches occur less frequently, constituting 32% of all switches. These switches typically happen between sentences or clauses, where one language is used in a complete sentence, and the next is initiated with the other language.

Example Sentences:

Intra-sentential: “ أنا رايح السوق،but I need to buy some groceries.” (Arabic followed by English within a sentence)

Inter-sentential: “ هو عنده مشكلة،and he can’t solve it.” (Arabic followed by English at the sentence boundary)

1. **Structural Features**

The structural analysis thus proceeds on syntactic patterns, namely, word order, agreement as well as the alignment of sentence elements. This structural analysis aims at identifying language-specific tendencies as well as how these are maintained or altered in language switching.

Findings: Word Order: Generally, it is maintaining Iraqi Arabic (VSO) and English (SVO) word orders, but there is a strikingly tendency for the English structure to prevail over sentences with intra-sentential switching; for instance, " أنا رايح السوق: VSO in Arabic" would switch to: "I need to buy some groceries": SVO in English, with that word order. Agreement between Subject and Verb: In many cases, Arabic being the subject of the sentence with an English verb, agreement concerning gender and number maintains according to Arabic syntax, while English would impose its own set of rules in syntax. There are some islands of exception, especially in informal context of code switching where agreement could be blurred.

Table (1) summery of structured features

Feature	Findings	Example
Switch points	68% intra sentential, 32% intra sentential	أنا رايح السوق. but I need to buy some ".groceries
Word order	Preference for SVO in English switches	| أنا رايح السوقneed to buy “.some bread
Agreement	Agreement maintained for Arabic subjects	يدرس، وهوbut he is working “.too
Verb non alignment	Tendency for English verbs to follow Arabic subjects	| أنا مشغول have a ".meeting now

3.6.2 Quantitative Evaluation of AI Performance

Performance of AI Parsers in Quantitative Assessment:

- Accuracy: This refers to the percentage of well parsed sentences over the total number of processed sentences.

- Precision: The fraction of correctly identified syntactic relationships out of total identified ones.

- Recall: Ratio of correctly identified relationships to total potential relationships.

- F1-Score: It is an overall measure of performance of model, calculated as the harmonic mean of precision and recall.

These results were compared with the baseline performance as represented by processing in a single language, using this in understanding the utility and success of AI adaptation for the mixed-language input context.

Evaluation Metrics:

Accuracy: Describes the overall proportion of sentences that have been parsed correctly.

Precision: Denotes the proportion of correctly identified syntactic relationships out of all identified relationships.

Recall: The proportion of all conditions in which the proper relationship could be identified**F1-Score**: Combines precision and recall into a single measure of performance.

Findings:

Table (2) The performance of the AI parser was evaluated on both the code-switched corpus and a baseline monolingual corpus (Arabic and English separately).

Metric	AI parsing (code-switched)	AI parsing (monolingual)
Accuracy	82%	94%
Precision	79%	93%
Recall	75%	92%
F1 – score	77%	92%

Analysis of Results:

Accuracy level: The AI parser works at an accuracy of 82% in parsing code-switched sentences, a performance that is lower than the 94% accuracy achieved while parsing monolingual sentences. This means that the parser is good at code-switching sentences but has a challenge with the mixed language trait of the data involved.

Precision level: The AI parser was able to identify syntactic relations, at a precision level of 79%, in code-switched sentences; however, this is slightly lower than its achievement for monolingual data (93%). This indicates that the parser had some issues with some complex syntactic interactions that arise in bilingual contexts.

Recall: The recall rate of 75% means that the AI parser can identify 75% possible syntactic relations in code-switched sentences, which is lower than the 92% for monolingual sentences. Lower recall may mean most of the relations were missed with a few exceptions, especially due to switching points. F1 Score: With an F1 score of 77% in code-switched sentences, it states a middle performance assessment where precision and recall were balanced against one another. In the case of monolingual data, the F1 score was 92%, signifying stronger performance overall in standard, monolingual inputs.

Discussion of Findings

The result of this analysis is very advantageous in terms of providing insights into the syntactic features of Iraqi Arabic-English code-switching and the efficiency of performance regarding the efficacy with which AI-based parsers handle the mixed languages. The qualitative analysis resulted in specific patterns found in switch points, word order, agreement, and verb-noun alignment. In contrast, the quantitative analysis put forward the strengths and challenges of AI-based parsers while processing mixed-language sentences. Among the findings of this study, 68% of the code-switching was found to be intra-sentential, where 32% were found to be inter-sentential. This resulted in a predominance of factors that are usually the same for all codeswitchers. In terms of both distribution and general usage tor bilinguals, intra-sentential switching is more prevalent than inter-sentential. This can be understood from the fact that, often when it occurs at an insert among different syntactic boundaries such as at noun phrases or verb phrases, the bilingual speaker keeps within a single frame the syntactic structures of both languages. Such would mean a high degree of syntactic flexibility because both languages can be seamlessly integrated by the bilingual speaker. Conversely, inter-sentential switches appear to be when the speaker moves from one thought or concept to another and uses different languages for emphasis or introduction to a new topic. Their lesser frequency of occurrence may indicate the tendency of speakers to use separate languages when conducting prolonged communication. Co-switch word order analysis clearly indicated a tendency for English head structures in code-switched sentences, irrespective of the structure governing the Arabic component. An Arabic verb usually assumes the VSO order, while an English verb resorts into SVO. But, it is clear from the code-switched sentences that, English word order dominates, indicating that once a bilingual draws English into his speech, he also adopts its syntactic framework in that respect. It exposes a greater role that English plays in the syntactic format of code-switched discourse, as bilinguals adjust on language use according to the functional relevance that a sentence bears. This shift might be taken to suggest that English is interpreted rather as a primary language in most interactions of bilinguals in informal contexts, like social sites or casual conversation. Subject-verb agreement in code-switches rather adhered to Arabic-grammar rules, even if the verb were in English. This is consistent with an earlier finding on rules for code-switching that such bilinguals embrace a lot of the syntax from the dominant or first language. It was observed that, with English verbs, there were still markers for that agreement, thus demonstrating the arabesque syntax even in meshed speech. But many of those agreements are not necessarily obeyed, especially in the informal context, thus showing that bilinguals are flexible with rules.Regarding verb-noun alignment, the study found that English verbs often followed Arabic subjects, which is a common pattern in code-switching between languages with differing syntactic structures. This alignment may reflect the influence of Arabic syntax, as the Arabic subject often precedes the verb in its native word order. Despite the English verb being used, this pattern of verb-noun alignment remained consistent, further demonstrating the influence of Arabic syntax on bilingual speech.

Conclusion

This research has provided a comprehensive exploration of the syntactic features of Iraqi Arabic-English code-switching and the evaluation of AI-based parsers in handling mixed-language data. By combining both qualitative and quantitative methods, this study has deepened our understanding of the intricate syntactic patterns that emerge when speakers blend two languages within the same discourse and has assessed the extent to which AI models can effectively parse such bilingual sentences. The qualitative analysis revealed key syntactic features of code-switching, including the predominance of intra-sentential switches and the influence of English word order on code-switched sentences. These findings highlight the complex linguistic behavior of bilingual speakers, who navigate between different grammatical systems while maintaining the fluidity and coherence of their communication. Additionally, the study demonstrated that while subject-verb agreement in code-switched sentences tends to follow Arabic rules, flexibility and adaptation are common in informal discourse, where agreement markers may be less strictly observed.

On the computational side, the AI-based parsers showed promising results in analyzing code-switched sentences, with an accuracy of 82%, although performance was lower compared to monolingual parsing. The AI parser struggled with the syntactic complexities introduced by bilingual sentences, particularly in terms of precision and recall. However, the overall performance, reflected in the F1-score, suggests that AI-based systems can be adapted to handle code-switching, though further improvements are needed to enhance their accuracy and effectiveness in mixed-language contexts.

References

- Al-Khatib, M. A., & Sabbah, E. (2008). Language choice in education: A case study of Jordanian public schools. International Journal of Bilingual Education and Bilingualism, 11(5), 568-586.
- Albirini, A. (2016). Modern Arabic sociolinguistics: Diglossia, variation, codeswitching, attitudes, and identity. Routledge.
- Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1, 844-857.
- Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT, 1, 4171-4186.

- Holes, C. (2004). Modern Arabic: Structures, functions, and varieties. Georgetown University Press.
- Jurafsky, D., & Martin, J. H. (2023). Speech and language processing. 3rd ed. Pearson.
- Jassem, Z. A. (2012). The sociolinguistics of English and Arabic in the Arab world. English Linguistics Research, 1(1), 148-171.
- Myers-Scotton, C. (1993). Social motivations for codeswitching: Evidence from Africa. Oxford University Press.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. Linguistics, 18(7-8), 581-618.
- Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.