Proceedings of the OSACT 2022 Workshop @LREC2022, pages 23–31
Marseille, 20 June 2022

AraSAS: The Open Source Arabic Semantic Tagger

Mahmoud El-Haj, Paul Rayson, Elvis de Souza*, Nouran Khallaf† and Nizar Habash‡

Lancaster University, UK
*Pontifical Catholic University of Rio de Janeiro, Brazil
†University of Leeds, UK
‡New York University Abu Dhabi, UAE
{m.el-haj, p.rayson}@lancaster.ac.uk,*elvis.desouza99@gmail.com,
†mlnak@leeds.ac.uk, ‡nizar.habash@nyu.edu

Abstract

This paper presents (AraSAS) the first open-source Arabic semantic analysis tagging system. AraSAS is a software framework that provides full semantic tagging of text written in Arabic. AraSAS is based on the UCREL Semantic Analysis System (USAS) which was first developed to semantically tag English text. Similarly to USAS, AraSAS uses a hierarchical semantic tag set that contains 21 major discourse fields and 232 fine-grained semantic field tags. The paper describes the creation, validation and evaluation of AraSAS. In addition, we demonstrate a first case study to illustrate the affordances of applying USAS and AraSAS semantic taggers on the Zayed University Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) (Palfreyman and Habash, 2022), where we show and compare the coverage of the two semantic taggers through running them on Arabic and English essays on different topics. The analysis expands to compare the taggers when run on texts in Arabic and English written by the same writer and texts written by male and by female students. Variables for comparison include frequency of use of particular semantic sub-domains, as well as the diversity of semantic elements within a text.

Arabic, English, Semantics, Corpus Linguistics, Taggers

1. Introduction

Semantic tagging is the process of associating an element of text data to a well-formed ontology or a lexicon (Rayson and Wilson, 1996; Rayson et al., 2004). While in other types of semantic annotation, the tagging can be applied to a whole text or to text fragments (e.g. sentences, words), in this paper we consider only the case of assigning labels (or tags) to words and multi-word expressions. The tags are assigned based on a pre-defined semantic lexicon indicating coarse-grained word senses. A lexicon refers to the component of a Natural Language Processing (NLP) system that contains semantic or grammatical information about individual words or word strings (Guthrie et al., 1996). This annotation can be considered as a tool for semantic enrichment of the text which facilitates the development of various types of NLP applications especially allowing a better performance for semantic search (Kogalovskii, 2018; Rayson et al., 2004). Moreover, semantic annotation is an important task in NLP, with the original semantic tagger being developed for English (Piao et al., 2015b; Piao et al., 2016b).

Unlike English and despite the increasing interest in research related to Arabic NLP, there is still a lack of well developed NLP tools and techniques that are required to advance the computational study or application of semantics. This is partly due to features of Arabic morphology and orthography which are very different to English and other Indo-European languages, as well as the lack of available corpus resources over time. Arabic is morphologically rich and complex (Habash, 2010; El-Haj and Rayson, 2016). In addition to its rich inflectional and derivational systems, Arabic has a large number of attachable clitics such as prepositions, and pronouns. Arabic orthography uses optional diacritical marks to indicate short vowels, and consonantal gemination. But these diacritics are almost never used beyond religious texts and children's books. The combination of rich morphology and underspecified orthography leads to a high degree of ambiguity: the Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009), e.g., returns 12 analyses per word on average. This ambiguity comes on top, and independently, of the kind of polysemy that is common in many languages (e.g. "set", "run" and "get" have multiple related meanings in English). For example the undiacritized Arabic word ولعين *wlʕyn*[1], returns four lemmas and 83 analyses using the CALIMAStar Arabic analyzer (Taji et al., 2018) inside of CamelTools (Obeid et al., 2020). The lemmas correspond to the *vocables*: لَعِين *laʕiyn* 'cursed' (adjective), عَين *ʕayn* 'Ain' (proper noun), وَلِع *waliʕ* 'passionnate' (adjective), and عَين *ʕayn* 'eye' (noun). The effect of morphological inflection and cliticization together with underspecified vowels can be demonstrated by contrasting two of these readings: وَلِعِينَ *waliʕiyna* 'passionnate [masc.pl]' and وَلِعَينٍ *wa+li+ʕaynĩ* 'and+for+an-eye'. When consider-

[1]Arabic Transliteration in the HSB Scheme (Habash and Rambow, 2005).

ing polysemy, the lemma عَين ʿ*ayn* by itself has around 50 meanings out of context (ibn Mukarram ibn Manzūr, 1290): besides 'eye', 'eighteenth letter of the Arabic alphabets', 'spy', 'envy', 'sun', 'rain' and 'water spring'. In this paper, we introduce the first open source Arabic Semantic tagger (AraSAS). Throughout the paper, we describe the process of creating, validating and evaluating AraSAS. In addition we analyse the semantic fields or domains of words used in ZAEBUC[2] corpus (Palfreyman and Habash, 2022; Habash and Palfreyman, 2022) which contains text written by bilingual students writings from different cities in the United Arab Emirates (UAE) who had just joined Zayed University. The text was created by undergraduate Arabic-English bilingual students as part of their degree, where the written assignment was to assess their language skills. The assignments written by the students formed the ZAEBUC corpus. The analysis provides semantic annotation to ZAEBUC as a new gold-standard language resource to increase the understanding of texts, especially through machine learning and NLP. In addition, the analysis helps in widening and supporting both comparative research and studies of the Arabic language from the perspective of English. Fuller details of the application of this tagger to ZAEBUC are presented in (Khallaf et al., 2022).

AraSAS is the Arabic equivalent to the well established English UCREL Semantic Analysis System (USAS) (Rayson et al., 2004). The USAS lexicon (Rayson et al., 2004) contains 21 major semantic fields (see Figure 1) with 232 sub-classes as the reference semantic ontology. USAS was used in the first prototype of AraSAS tagger (Mohamed et al., 2013), which was never been publicly released. The authors used the Buckwalter Arabic Morphological Analyser (Buckwalter, 2004) to compile a list of Arabic lemmas. Buckwalter also provides English glosses (equivalent translations) of those lemmas. The English translations were used to match against the entries in the USAS English lexicon, where they then compiled a lexicon entry for each Arabic lemma containing the union of all tags from the entries of all its possible equivalents. Although the authors managed to match 71% of the lemmas to the USAS lexicon, the process itself was error prone due to the out-of-context matching process (Zawahreh, 2013). The process resulted in a lexicon containing 37,312 lemmas. A post-editing process was performed on just 4% of the total lemmas. The lemmas were sorted by lemma frequencies in the Leeds Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell, 2006), with the post-editing being focused on the most frequent lemmas in order to maximise coverage of typical texts. Our new release of AraSAS provides a much extended and edited semantic lexicon and an open source semantic tagger tool that is developed for the Arabic language. AraSAS has been integrated with CAMeL Tools[3] to provide a Python platform for researchers working on Arabic NLP (Obeid et al., 2020).

[2] http://www.zaebuc.org

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Figure 1: USAS Tagger Major Discourse Fields

Both AraSAS and USAS will help in analysing the semantic domains of words used in ZAEBUC by the bilingual university-level students in Arabic and English. The taggers will help in annotating each word in those texts by giving each word a semantic domain/sub-domain (e.g. "B2: Health and disease" or "A6.1 Comparing: Similar/different")[4].

2. Related Work

Semantic tagging is an umbrella term for a wide variety of other terms and tasks related to linguistic annotation of corpora. These can include mark up tasks such as Named Entity Recognition (locations, names, dates, times and organisations), semantic role labelling (goals, agents and results), word sense disambiguation (fine-grained dictionary senses), summarisation (reducing the length of a text while retaining its core meaning), or sentiment analysis (annotation for positive and negative opinions about a product or service).

A core task implemented by the UCREL Semantic Annotation System (USAS) (Rayson et al., 2004), is to assign coarse-grained semantic fields to all words and phrases in a text. Originally applied for content analysis of market research interview transcripts, the USAS tagger is a knowledge based system incorporating manually curated semantic information in large single word and multi-word expression (MWE) lexicons, approximately 80,000 words and MWEs in total. The job of the tagger is then to select the contextually appropriate semantic tag to best represent the broad semantic field of a word or MWE. The semantic tags are taken from a

[3] CAMeLTools: An Open Source Python Toolkit for Arabic Natural Language Processing which offers a robust Arabic morphological analysis, dialect identification, named entity recognition and sentiment analysis.

[4] The categories are based on a hierarchical semantic tag set. The first letter shows the major discourse field as shown in Figure 1 (e.g. B refers to 'the body and the individual', while the 6 in B6 refers to the sub-category 'health and disease'.

Word	Gloss	POS	Lemma	Semantic Tags
تشهد *tšhd*	witnesses	verb	1_شَهِد *šahid*	S9 A10+@ X3.4 G2.1 Q2.1 S7.1-@ X3.2
دولة *dwlħ*	state	noun	1_دَوْلَة *dwlħ*	G1.1c W3 F4/M7 M7
الإمارات *AlĂmArAt*	Emirates	noun	1_إِمَارَة *ĂimAraħ*	Q1.1 A6.2+ X4.1 Q1.2 O4.1 S9 B2
العربية *Alςrbyħ*	Arab	adj	1_عَرَبِيّ *ςrabiy∼*	Z2/Q3
المتحدة *AlmtHdħ*	United	adj	1_مُتَّحِد *mut∼aHid*	S5+ A1.1.1
تطورا *tTwrA*	development	noun	1_تَطَوُّر *taTaw∼ur*	A5.1+/A2.1 T2++ A2.1+ H1 A3+/A11.1
كبيرا *kbyrA*	great	adj	1_كَبِير *kabiyr*	N3.2+ N5+ A11.1+ A5.1+ X5.2+ A13
.	.	punc	.	PUNC

Table 1: Example of tagging an Arabic sentence (POS: part-of-speech tag).

taxonomy of 232 labels grouping together word senses that are connected to the same topic, e.g. the 'education' tag P1 is assigned to words and MWEs such as 'academic', 'coaching', 'coursework', 'deputy head', 'exams', 'PhD', 'playschool', and 'revision notes'. The English tagger performs this task at around 91% accuracy (Rayson et al., 2004).

Using a similar knowledge-based model with manually created lexicons, semantic taggers for other languages were created e.g. Finnish (Löfberg, 2017) and Russian (Mudraya et al., 2006). More recently, bootstrapping approaches have been evaluated to more quickly generate prototype semantic lexicons in new languages (Piao et al., 2015a), alongside crowd-sourcing methods to see whether non-expert native speakers could assist in the creation and checking of such resources (El-Haj et al., 2017). This has resulted in a proliferation of semantic taggers in multiple languages (Piao et al., 2016a), as well as applying further contextual disambiguation methods to apply more fine-grained taxonomies in a historical context (Piao et al., 2017), and multi-task machine learning methods to derive annotation knowledge from manually tagged corpora and pre-trained embeddings (Ezeani et al., 2019). For a more detailed discussion of the development of USAS, see (Löfberg and Rayson, 2019).

Other than the previous work on AraSAS, most research on Arabic semantic annotation is limited to semantic role labelling (Al-hadi et al., 2016). Similarly, semantically annotated corpora and other tools for Arabic are still in the early stages of creation, with very few available resources (Saleh and Al-Khalifa, 2009).

3. AraSAS Semantic Tagger

The new Arabic Semantic Annotation System (AraSAS) was developed in $Python$ 3 and makes use of several other Python packages in its pipeline. The AraSAS pipeline to transform raw Arabic text into semantically tagged output is illustrated in Figure 2.

As shown in Figure 2, the first part of the pipeline is sentence segmentation, which is performed using the Natural Language Toolkit (NLTK) (Loper and Bird, 2002). For the segmentation to work properly for Arabic texts, we also needed to replace right-to-left with left-to-right punctuation marks (e.g. the Arabic question mark '؟' to the English/Latin question mark '?').

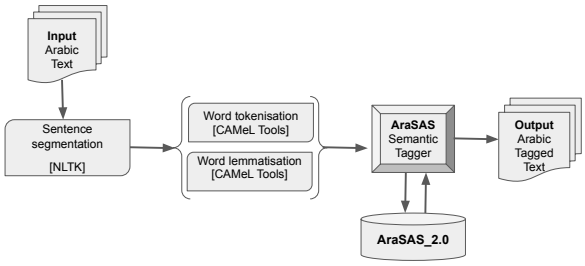


Figure 2: AraSAS pipeline

Once sentences are identified, AraSAS calls CAMeL Tools (Obeid et al., 2020) to tokenise the sentences into tokens. Those tokens are then disambiguated using a morphological analyser, which ranks the most probable analysis for a word based on its lemmatisation and part of speech annotation.

As an example we used AraSAS to tag the following Arabic sentence from ZAEBUC: تشهد دولة الامارات العربية المتحدة تطورا كبيرا 'The United Arab Emirates is witnessing a great development.' The result of the semantically tagged sentence is shown in Table 1, where each token is displayed in a new line and each semantic tag is separated by a white space. Besides the use of a lexicon, a few regular expressions are also applied to finding punctuation and numbers, both receiving their own distinct semantic tags.

We have made AraSAS freely available open source for academic use[5]. AraSAS is also available as a web-tool, where users can type in or paste their text to be tagged[6].

3.1. AraSAS Lexicon Creation

We used the first draft of the AraSAS lexicon (henceforth, `AraSAS_1.0`) (Mohamed et al., 2013) to cre-

[5] `https://github.com/UCREL/AraSAS`
[6] `http://ucrel-api.lancaster.ac.uk/`

	Lemma	Translation	POS	Semantic Tag
AraSAS_1.0	عَلِيّ ʕ*aliy*~	supreme, high, Ali	–	A1.1.1 A5.1+++ Z1m
AraSAS_2.0	عَلِيّ ʕ*aliy*~	supreme, high	adjective	A1.1.1 A5.1+++
	عَلِيّ ʕ*aliy*~	Ali	proper noun	Z1m

Table 2: Lemma representation in both original and CAMeL list annotated with Semantic tags

ate a verified lexicon that we can use with the newly created AraSAS semantic tagger (henceforth, AraSAS_2.0).

One of the main shortcomings of the AraSAS_1.0 lexicon is that it used a reduced basic representation of Arabic lemmas. In contrast, the CAMeL database, which is based on BAMA/SAMA, provides number markings for different meanings of a lemma. For example, the lemma عَـلِيّ ʕ*aliy*~ has two variants: ʕ*aliy*~_*1* 'supreme;high (adj)', and ʕ*aliy*~_*2* 'Ali (proper noun)'. These two variants are collapsed into one lemma in AraSAS_1.0. Most of these number ids overlap with POS distinctions as in the above example. In the CAMeL database, there are 42,226 (lemmas with ids and POS), corresponding to 37,613 basic lemmas (no ids), and 40,795 basic lemmas with POS. We keep the lemma disambiguation markers and POS while building the AraSAS_2.0 lexicon as much as possible in an attempt to increase the precision of the semantic tagging process. Originally, when creating the English USAS lexicon, words with their POS tags were used for lexicon entries, and gradually this was updated to include lemmas with POS once a lemmatiser was included in the English processing pipeline.

We began by creating a list of the most frequent lemmas in the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) and cross matched it to the lemmas in AraSAS_1.0. The PATB lemma-list was created by selecting 32,000 words from the full PATB words list, which was done to match the number of Arabic words in ZAEBUC (212 documents with each containing an average of 160 Arabic words). We then lemmatised the words and normalised digits using CAMeL tools, which resulted in a PATB lemma-list of 4,500 lemmas. Matching AraSAS_1.0 to PATB lemma-list we ended up with 200 new lemmas, which were found in PATB lemma-list but not in AraSAS_1.0. We then combined the AraSAS_1.0 lemmas with the new 200 lemmas to create an updated list of lemmas as in Figure 3.

We asked a linguist, who is also the fourth author of this paper, to manually check the validity of given semantic tags for a random sample of 150 lemmas from AraSAS_1.0. The linguist found that a large number of the lemmas were assigned unrelated semantic tags. This was mainly due to the fact that a single word may have multiple meanings. This was not accounted for in AraSAS_1.0 where the authors compiled a lexicon entry for each Arabic lemma containing the union of all tags from the entries of all its possible equivalents

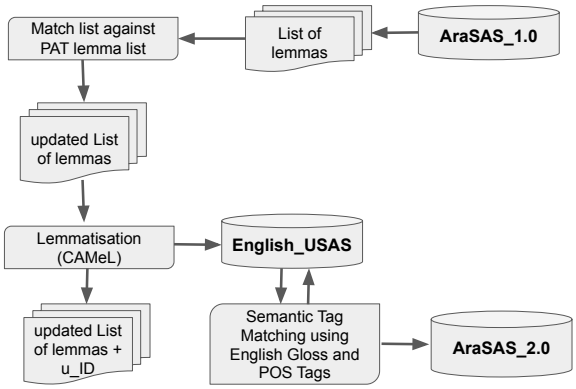


Figure 3: The process of creating AraSAS_2.0 lexicon

in English. Although this would result in a higher coverage, it reduced precision and accuracy (Mohamed et al., 2013). To overcome this problem we decided to use CAMeL tools lemmatiser and part of speech (POS) tagger to help capture the different meanings of all lemmas in the updated list of lemmas as illustrated in Figure 3. For a single lemma with different possible meanings, the CAMeL lemmatiser follows each of those meanings with a unique number (u_ID) to help differentiate between them (e.g. work_1, work_2) (Obeid et al., 2020). The process resulted in the *Updated List of Lemmas + u_ID* as shown in Figure 3.

To sustain the different meanings, we included u_ID in a new column in the lexicon to help identify such differences as shown in Table 2. As with the Buckwalter Morphological Analyser, CAMeL also provides the English glosses (equivalent translations) of those lemmas, which we used to match against the USAS English lexicon in order to produce a more accurate AraSAS lexicon (AraSAS_2.0).

The linguist validated the same 150 lemmas as in AraSAS_1.0 by comparing to AraSAS_2.0, but this time considering the different meanings for a single lemma using u_ID. The linguist found that the lemmas from AraSAS_2.0 provided a clear division between the senses, since each entry with a different meaning was treated as a different lemma.

The second step after that was to match AraSAS_2.0 entries against USAS English lexicon using POS tags and English gloss for each lemma. The process resulted in an addition of 4,260 entries on top of the original 37,312 entries found in AraSAS_1.0 (the new 200

lemmas from the Penn Arabic Treebank are included in the 4,260 lemmas). The reason for this increase in AraSAS_2.0 was due to the use of u_ID, which resulted in a total of 41,572 entries.

3.2. Validating AraSAS_2.0

AraSAS_1.0 was based on lemmas appearing in the Leeds Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell, 2006). This lexicon required some modifications in order to maximise its coverage, as well as normalise the lemmas to be compatible with CAMeL Tools (Obeid et al., 2020). Firstly, we compared the semantic tags attached to the first one thousand lemmas in the AraSAS_1.0 list with AraSAS_2.0. Areas where significant differences have been found include both lemmatisation and semantic tags. For instance, as represented in Table 2 the two variants of the lemma عَلِيّ *ʿaliy~* with their different POS tags were collapsed in AraSAS_1.0, but are distinguished in AraSAS_2.0. This lemma division has a high impact on the semantic annotation process. On the other hand, in AraSAS_1.0 there was some disagreement upon some annotated semantic tags such as in lemma مع *mς* 'with/together' which was wrongly annotated with T1.1.2 (Referring to Time: General: Present; simultaneous), we believe there are no cases for time correlated with this Arabic lemma. However, this meaning is correlated with the English word 'together', which refers to 'at the same time'. For this reason, a manual modification of the semantic tags for the first most frequent 1000 Arabic lemmas involved removing or adding a semantic sense and rearranging the previously annotated senses according to the new added lemmas.

Secondly, the process of building AraSAS_2.0 was challenging in terms of matching and aligning the lemmas to AraSAS_1.0. In this case we matched both lists using the English translation provided in AraSAS_1.0 and the English gloss provided by the CAMeL analysis along with the POS tags to produce the new merged list. Moreover, in AraSAS_1.0 common orthographic inconsistencies, e.g., among various Alif-Hamza forms: اأإآ *AÂĂĀ*, resulted in mismatches between the two lists.

Matching using the English translations and glosses resulted in adding 286 new untagged lemmas that were not analysed by CAMeL POS tagger. Starting with manual analysis and semantically tagging these lemmas, we found they are usually due to transliteration and misspellings. Examples include (a) English words written with Arabic letters such as هاوس *hāwis* 'house' and تيليفيجن *tīlīfijan* 'television'; (b) English proper nouns written in Arabic letters such as رولاند *Rūlāand* 'Roland'; and (c) misspelled words such as اوكسترا *Ūksitrā* rather than اورکسترا *Ūrksitrā* 'orchestra'.

This validation process was followed by manually analysing the 10,023 lemmas assigned the Z99 semantic tag, which is used when there is no match (unmatched category) to any of the tags found in AraSAS_2.0 tagging list. We manually annotated the 1,300 most frequent PATB lemmas in order to prioritise our efforts in assigning a tag other than Z99 as widely as possible. Around 600 lemmas of the manually checked 1,300 Z99 lemmas were found to be Personal-Names (Z1), Geographic-Names (Z2) or Other-Proper-Names (Z3).

3.3. AraSAS Evaluation

We evaluated AraSAS lexical coverage by tagging two different sets of texts in Arabic: one from Arabic blogs (composed of 1,114,535 tokens, according to CAMeL Tools tokeniser) and another from Arabic newspapers (1,108,058 tokens).

Running the lexical coverage experiment, we found that AraSAS lexical coverage ranges from 96% of tokens in blogs texts to 96.8% in news texts. The results shows that AraSAS_2.0 to have a high lexical coverage.

Additionally, using the ZAEBUC dataset (described in Section 4.), we evaluated the proportion of untagged words from the English sub-corpus (annotated by the English original USAS) and the Arabic sub-corpus (annotated by the recently developed AraSAS_2.0). The English tagger showed a lexical coverage of 99.3%, while applying AraSAS resulted in coverage of 98.3%. For future work we will work on manually tagging a larger set of AraSAS entries and calculating recall, precision and F-measure scores to assess the tagging quality.

Originally, English USAS as well as USAS in other several languages, have been assessed and evaluated to measure tagging quality. The work her is a release of the AraSAS semantic tagger tool, which we believe is going to be useful for researchers working on Arabic NLP and Arabic semantics. In previous, work we report the process used to evaluate the quality of lexicon bootstrapping for different languages, including Arabic, as shown in (El-Haj et al., 2017) and (Piao et al., 2016b).

4. ZAEBUC Corpus

Zayed Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) is composed of bilingual students writings from different cities in the United Arab Emirates (UAE) who just joined Zayed University. The students were asked to do two assignments to assess their language skills, one in Arabic and another in English. Their essays, that are part of the assignment, were collected to compose the bilingual writers corpus of ZAEBUC.

Students could choose from three different topics and they did not have to choose the same topic for the writings in Arabic and English, although most of them did. The resulting ZAEBUC corpus is not balanced– there are more female than male authors, more English than Arabic essays, and most of the students chose the 'social media' topic over 'development' and 'tolerance'.

Despite the differences, we were still able to establish fruitful comparisons by looking at their writings in terms of semantic domains.
The students essays were manually validated by academics who detected spelling mistakes replaced them by the corrected words for the sake of the experiments described in this paper, as otherwise the semantic tagging would be less accurate.

5. Experimental Work

As mentioned earlier, Arabic texts from ZAEBUC were semantically annotated using AraSAS, while the English ones were tagged by the original English USAS. Both systems share the same tagset and tagging methodology. The ZAEBUC dataset is unbalanced, it is composed of 603 essays,[7] of which 215 were written in Arabic and 388 in English. The coverage percentage for each language is shown in Table 3. Tokens not tagged by AraSAS were later manually annotated and added into the lexicon to improve its coverage.

	Arabic	**English**
Texts	215	388
Tokens	34,442	97,994
Tagged	33,887	97,354
Untagged	555	640
Coverage	98.4%	99.3%

Table 3: ZAEBUC Composition

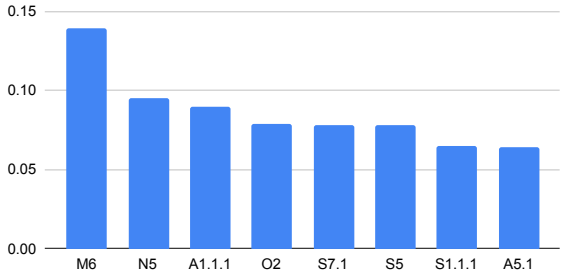
To compare the AraSAS and USAS we a modified version of the Type Token Ratio (TTR) formula (Richards, 1987), where instead we consider the total number of unique semantic tags rather than words (types). We call the updated formula the Semantic Token Ratio (STR) as shown in the following formula:

$$STR = \frac{N_{tag}}{T} \quad (1)$$

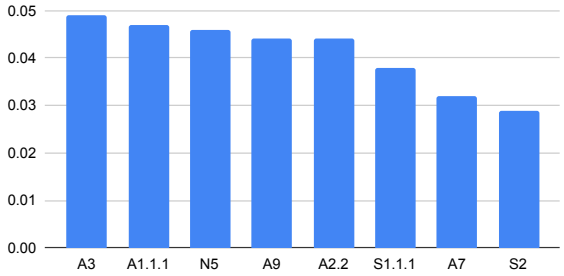
where STR is the Semantic Token Ratio, N_{tag} is the number of tokens that received a given semantic tag, and T is the total number of tokens in the sub-corpus, allowing the comparison of unbalanced sub-corpora like the ones featured in ZAEBUC.

5.1. Comparing Texts from Bilingual Authors

As stated earlier, the authors from ZAEBUC are Arabic-English bilingual speakers, but not all of them wrote in both Arabic and English as shown in the corpus description in Table 3. The number of texts in the English sub-corpus exceeds the number of Arabic texts by around 80%. As shown in the table, the number of tokens is far more unbalanced than the texts, there are 184% more tokens in English than in Arabic. This resulted in an average of 252 tokens for each text in English, while Arabic texts have an average of 160 tokens per article due to the high use of clitics and affixes/suffixes in Arabic (García-Barrero et al., 2013)
Figure 4 shows the eight most frequent semantic tags in texts written in Arabic and English after running AraSAS and USAS. Semantic tags related to punctuation and grammar are not featured as otherwise they would represent most of the relative occurrences while not being relevant to the discussion.
The Y axis in Figure 4 shows the percentage of the sample that the semantic tag occupies. For example, the most frequent semantic tag in texts written in Arabic is `M6` (Location and direction), which represents 13.7% of all tokens in the Arabic sub-corpus, while the most frequent semantic tag in English is `A3` (Being)[8].



(a) Texts in Arabic



(b) Texts in English

Figure 4: Semantic Type Ratio across languages

It is worth noting that tags in Arabic seem to be more skewed than English ones. Figure 4 shows that around 68% of the tokens in Arabic are distributed among the eight most frequent semantic tags, while for English the eight most frequent ones represent only around 33% of the sub-corpus. One explanation for it lies in the fact that we are not displaying in the figure tags that are grammatical – `Z5` (Grammatical bin) represents 33% in English and only 25% in Arabic, while `Z8` (Pronouns etc.) represents 11% in English and only 7% in Arabic. The semantic tag `M6` (Location and direction), while the most used in Arabic, is not be found in the eight most frequent tags in the English sub-corpus. At the

[7] These numbers are based on an early release of the ZAEBUC Corpus v0.1.

[8] Full list of tags used by USAS and AraSAS: https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

Tag	Label	MFT	STR
Z5	Grammatical bin	.	0.25
M6	Location and direction	في *fy* 'in/inside	0.13
N5	Quantities	فرد *frd* 'one/individual	0.09
A1.1.1	General actions, making	موقع *mwqς* 'location'	0.08
O2	Objects generally	رد *rd* 'to reply'	0.07
S7.1	Power, organizing	كبير *kbyr* 'large'	0.07
S5	Groups and affiliation	مجتمع *mjtmς* 'community'	0.07
Z8	Pronouns etc.	أنّ *Ân* 'that'	0.07
S1.1.1	General	اجتماعي AjtmAςy 'social'	0.06
A5.1	Evaluation: Good/bad	سلبيّ *slby* 'negative'	0.06

Table 4: Semantic tags in Arabic texts (10 out of 222 tags) *MFT: Most Frequent Token, STR: Semantic Type Ratio*

same time, the fact that A3 (Being) is the most frequent tag for texts in English is likely due to the grammatical role of the verb "to be", making it not as much as frequent in Arabic. Interestingly, the semantic tags N5 (Quantities), A1.1.1 (General actions, making, etc.) and S1.1.1 (General) are frequent in both languages, although always mostly more used in Arabic.

The English sub-corpus showed particular preference for A9 (Getting & giving; possession), A2.2 (Affect: Cause/Connected), A7 (Definite (+ modals)) and S2 (People), while Arabic show more usage other semantic tags such as, O2 (Objects generally), S7.1 (Power, organising), S5 (Groups and affiliation) and A5.1 (Evaluation: Good/bad). Tables 4 and 5 show the 10 most frequent semantic tags in the Arabic and English sub-corpora and including the Z tagset subcategories Z5 Z8 that are used to refer to grammatical items and pronouns.

Tag	Label	MFT	STR
Z5	Grammatical bin	the	0.33
Z8	Pronouns etc.	it	0.11
A3	Being	be	0.04
A1.1.1	General actions, making	way	0.04
N5	Quantities	many	0.04
A9	Getting&giving; possession	have	0.04
A2.2	Affect: Cause/Connected	have	0.04
S1.1.1	General	social	0.03
A7	Definite (+ modals)	can	0.03
S2	People	people	0.03

Table 5: Semantic tags in English texts (10 out of 210 tags) *MFT: Most Frequent Token, STR: Semantic Type Ratio*

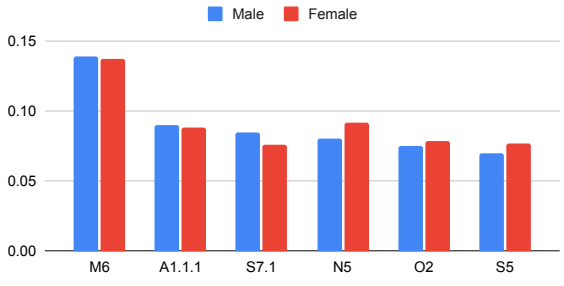
5.2. Comparing Texts Written by Male and Female Authors

We are also able to compare ZAEBUC texts by their authors' gender so we can assess any possible gender bias in the students' writings. The number of texts for each gender is extremely unbalanced, as seen in Table 6 show the distribution of authors by gender, where 90% of the students identified themselves as females.

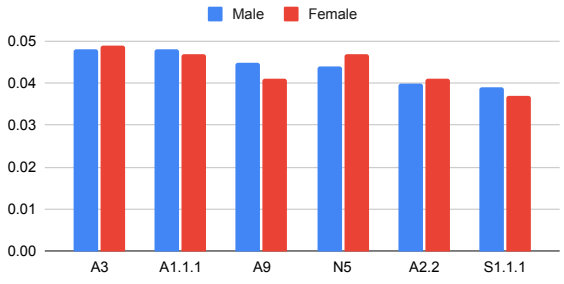
	Texts	Tokens
Female (Arabic)	199	32,115
Female (English)	344	87,804
Male (Arabic)	16	2,327
Male (English)	44	10,190

Table 6: Distribution authors by gender

Figure 5 shows the six most frequent semantic tags in texts from male and female authors in both languages. Once again, semantic tags related to punctuation and strictly grammatical ones are not shown.



(a) Texts in Arabic



(b) Texts in English

Figure 5: Semantic Type Ratio across authors' gender

The Y axis in the figure shows the percentage of the sample that the semantic tag occupies. The most frequent semantic tag in texts from authors from both genders in Arabic is M6 (Location and direction), representing 13.65% for the female writers and 13.92% for male writers. As for the English texts, both genders used A3 (Being) the most, as expected and explained when we compared languages in Section 5.1.. Texts by female writers show more frequent use of the semantic domain of N5 (Quantities) than texts by male writers, while A1.1.1 (General actions, making etc.) was used in a comparable proportion in both but more frequently in texts by male writers. Additionally, O2 (Objects generally) and S5 (Groups and affiliation) were more frequently used by female writers, although the semantic domain of S7.1 (Power, organizing) was used more frequently by male writers than female writers.

When it comes to English texts, other relevant semantic tags are N5 (Quantities) and A2.2 (Affects: Cause/Connected), it was found that female writers used those tags more than the male writers, while A1.1.1 (General actions, making etc.) was more applied by males, along with A9 (Getting & giving; possession) and S1.1.1 (General).

6. Conclusion and Future work

In this paper, we have presented the first open source Arabic Semantic Tagger (AraSAS). Building on prior work, we have significantly improved and extended the semantic lexicon which forms the linguistic knowledge base on which the AraSAS tagger relies. We have also created a new software tool in Python, which uses NLTK for sentence segmentation, and CAMeL Tools for tokenisation, and morphological analysis in terms of lemmatisation and POS tagging. The semantic taxonomy applied to words and multi-word expressions is the same as that used in the English and other language semantic taggers meaning that cross-lingual comparisons become possible at the level of coarse-grained semantic fields. We have made AraSAS freely available open source for academic use[9]. AraSAS is also available as a web-tool, where users can type in or paste their text to be tagged[10].

In terms of evaluation, we first considered the coverage of the semantic lexicon in terms of how many tokens in a corpus are present and matched in the lexicon. Very good coverage figures were obtained, 96% of tokens in blogs and 96.8% in news texts, which are comparable to the English USAS tagger. In addition, we performed a number of experiments on a corpus of Arabic-English writing (603 essays) by bilingual students in the UAE. Using AraSAS facilitates comparison of concepts and topics across the two languages in general, and to compare texts written by male and female authors. This serves to illustrate just the beginnings of the analysis possibilities that having comparable semantic tagging systems in two or more languages.

In terms of future work, we will focus on extending the single word lexicon along with a lexicon of multi-word expression patterns, and develop and apply further methods for bootstrapping the coverage, e.g., using word vectors, and for disambiguation, e.g., by applying deep learning methods.

[9] https://github.com/UCREL/AraSAS

[10] http://ucrel-api.lancaster.ac.uk/

7. Bibliographical References

Al-hadi, M., Ba-Alwi, F., Al-Baltah, I., Sana'a, Y., and Al-Gaphari, G. (2016). Survey of semantic annotation of arabic text. In *1st Scientific Conference on Information Technology and Networks(SCITN'2016)*.

Al-Sulaiti, L. and Atwell, E. S. (2006). The design of a corpus of contemporary arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

El-Haj, M. and Rayson, P. (2016). Osman—a novel arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255.

El-Haj, M., Rayson, P., Piao, S., and Wattam, S. (2017). Creating and validating multilingual semantic representations for six languages: Expert versus non-expert crowds. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 61–71, Valencia, Spain, April. Association for Computational Linguistics.

Ezeani, I., Piao, S., Neale, S., Rayson, P., and Knight, D. (2019). Leveraging pre-trained embeddings for Welsh taggers. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 270–280, Florence, Italy, August. Association for Computational Linguistics.

García-Barrero, D., Feria, M., and Turell, M. T. (2013). Using function words and punctuation marks in arabic forensic authorship attribution. In *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*, pages 42–56.

Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S., and Buckwalter, T. (2009). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B. M. (1996). The role of lexicons in natural language processing. *Communications of the ACM*, 39(1):63–72.

Habash, N. and Palfreyman, D. (2022). ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings*

of the Conference of the Association for Computational Linguistics (ACL), pages 573–580, Ann Arbor, Michigan.

Habash, N. Y. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

ibn Mukarram ibn Manzūr, M. (1290). Lisan al-arab.

Khallaf, N., de Souza, E., El-Haj, M., and Rayson, P. (2022). Semantic domains across topics, genders and languages.

Kogalovskii, M. (2018). Semantic annotating of text documents: Basic concepts and taxonomic approach. *Automatic Documentation and Mathematical Linguistics*, 52(3):134–141.

Löfberg, L. and Rayson, P., (2019). *Developing multilingual automatic semantic annotation systems*, pages 94–109. Cambridge University Press, June.

Löfberg, L. (2017). *Creating large semantic lexical resources for the Finnish language*. Ph.D. thesis, Lancaster University.

Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Maamouri, M., Bies, A., Buckwalter, T., and Mekki, W. (2004). The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Mohamed, G., Hardie, A., and Potts, A. (2013). Arasas: a semantic tagger for arabic. In *Second Workshop on Arabic Corpus Linguistics*.

Mudraya, O., Babych, B., Piao, S., Rayson, P., and Wilson, A. (2006). Developing a russian semantic tagger for automatic semantic annotation, October. Also published in Russian (pp. 282-289); Corpus Linguistics 2006 ; Conference date: 10-10-2006 Through 14-10-2006.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032.

Palfreyman, D. and Habash, N. (2022). Zayed arabic-english bilingual undergraduate corpus (zaebuc).

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., and Rayson, P. (2015a). Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274, Denver, Colorado, May–June. Association for Computational Linguistics.

Piao, S. S., Bianchi, F., Dayrell, C., D'egidio, A., and Rayson, P. (2015b). Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., and Mudraya, O. (2016a). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2614–2619, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Piao, S. S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., et al. (2016b). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2614–2619.

Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., and Alexander, M. (2017). A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language*, 46:113–135, November.

Rayson, P. and Wilson, A. (1996). The acamrit semantic tagging system: progress report. In *Language engineering for document analysis and recognition, LEDAR, AISB96 workshop proceedings*, pages 13–20.

Rayson, P., Archer, D., Piao, S., and McEnery, A. M. (2004). The ucrel semantic analysis system.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Saleh, L. M. B. and Al-Khalifa, H. S. (2009). Aratation: an arabic semantic annotation tool. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, pages 447–451.

Taji, D., Khalifa, S., Obeid, O., Eryani, F., and Habash, N. (2018). An Arabic morphological analyzer and generator with copious features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140–150, Brussels, Belgium, October. Association for Computational Linguistics.

Zawahreh, F. A. S. (2013). A linguistic contrastive analysis case study: Out of context translation of arabic adjectives into english in efl classroom. *International Journal of Academic Research in Business and Social Sciences*, 3(2):427.