Idiomatic MWEs and Machine Translation
A Retrieval and Representation Model: the AraMWE Project

Giuliano Lancioni[1]
Roma Tre University
lancioni@uniroma3.it

Marco Boella[1]
University of Rome "La Sapienza"
marco.boella@alice.it

Abstract

A preliminary implementation of AraMWE, a hybrid project that includes a statistical component and a CCG symbolic component to extract and treat MWEs and idioms in Arabic and English parallel texts is presented, together with a general sketch of the system, a thorough description of the statistical component and a proof of concept of the CCG component.

1 Introduction

We present the AraMWE Project[2], a hybrid model able to identify and represent Idiomatic Multi-Word Expressions (IMWE) in Arabic texts. Firstly IMWEs are identified in texts through standard computational quantitative-statistic strategies independent from linguistic knowledge. Then, a formal grammar theory, namely Combinatory Categorial Grammar (CCG), helps to parse and represent the IMWE structure, in order to improve recognition/generation in machine and machine-assisted translation and automatic alignment of specific elements in multilingual texts.

Chapter 2 presents some definitions on IMWE, CCG, Translation Memories and alignment, with related glance on current trends of research. In Chapter 3 the working model and the process flow of AraMWE project will be described, with a special focus on automatic recognition of given IMWE patterns and the strategies we adopted to account for IMWEs in a CCG environment. Chapter 4 gives information on data used and model testing and evaluation, and Chapter 5 closes the paper with some conclusions and an outlook on future developments.

[1] This paper is the result of joint work. However, the authorship can be attributed as follows: 1, 2.1, 2,2, 3.1 and 4 have been written by Boella, 2.3, 3.2 and 5 by Lancioni.
[2] host.uniroma3.it/docenti/lancioni/AraMWE.

2 Subject definitions and related research

2.1 Idiomatic Multiword Expressions

Multi-Word Expressions (MWE) are usually identified in literature with sequences of two or more words that have stronger relationships among themselves rather than with other sentence elements (Cacciari and Tabossi, 1993) or, following another definition, "a multiword unit or a collocation of words that co-occur together statistically more than chance" (Hawwari et al., 2012:24).

Studies on MWEs tend to suggest fluid and smooth classification criteria, which overlap with each other and form a continuum rather than defining sharp subsets (Sag et al., 2002).

The first parameter is semantic in nature and concerns compositionality. On the lower side we find MWEs whose meaning can be guessed by "composing" the meaning of the single elements (e.g. *the president of the republic*). Other MWEs have a medium degree of compositionality i.e. the resulting meaning is not merely a sum of that of the single elements, but somehow still related (e.g. *to carry coals to Newcastle*, which means 'to do something pointless'), up to those MWEs whose meaning has nothing to do with the single elements e.g. the often cited *to kick the bucket* 'to die', or *to spill the beans* 'to reveal secrets' (Cacciari and Tabossi, 1993).

The other main parameter involves morphosyntax: each element occurring in a MWE has a different degree of flexibility, in terms of position (MWE can contain non-MWE elements) and inflection (verb conjugation and noun declension).

Beside criteria of composition and flexibility, MWEs can be further classified according to the main parts of speech involved, e.g. Noun + Noun (NN), Verb + Noun (VN), Verb + Preposition (VP) and so on. These classes seem to have a certain rate of homogeneous behavior involving compositionality and flexibility, e.g. NN seem to be more composi-

tional and less flexible than VN (Cacciari and Tabossi, 1993).

These assumptions clearly don't set clear boundaries and show how difficult it is trying to define which MWE can be fully recognized as idiomatic. Idiomaticity seems obviously connected with low compositionality and relatively low flexibility (in positioning especially), but a clear definition is far to be outlined (Pawley, 1983), even if a long tradition of studies assigns to "idiomaticity" just the same meaning of "compositionality" (see for example Diab and Bhutada, 2009). For the purposes of our work, we provisionally call Idiomatic MWEs those multi-word expressions semantically non-compositional and syntactically non-conforming (see also Kavka and Zybert, 2004).

Related work: Concerning NLP approach to MWEs in Arabic, recent studies focus on two main directions, the construction of annotated repositories of MWEs and the automatic detection and extraction of MWEs from texts. Approaches for the first issue vary from those fully unsupervised (Cook et al., 2007) to more recent hybrid models that include supervised procedures to improve size and correctness of the list (Hawwari et al., 2012; Diab and Bhutada, 2009). Several works concern instead the automatic extraction of MWEs, with strong statistical approaches (Al Khatib and Badarneh, 2010; Moirón et al. 2006). Other recent models focus on parallel strategies to feed models with linguistic or statistical information needed to discern MWEs, especially for nominal ones (N+N) (Attia et al., 2010).

2.2 Translation memories and alignment

In the field of machine-assisted translation the collections of bilingual texts known as Translation Memories (TMs) aid human translator by providing sentences or larger text chunks in a given language, together with the 'aligned translation in another language, or other languages. Through strict or fuzzy search a translator can look up in the TM for the best match for the word context needed to perform correct translation. The employ of TMs is mainly as commercial and professional tool, and TM implication in computational and corpus linguistics was scarcely investigated, nevertheless some recent studies aim to reduce the size of aligned text chunks by using parsing systems, from sentences to sub-sentence elements, with the goal to get a complete aligned, cross-referenced bilingual parallel corpus (Lagoudaki, 2006).

Related Work: Many studies propose models to deepen TM alignment, in order to pair not only paragraphs and sentences, but also phrases, words and even word constituents (Simard, 2003). Among works that treat TMs specific to less represented languages focusing on an unsupervised approach, Chuang et al. (2005) show how to build a Chinese-English TM integrating statistical and linguistic information, and trying to analyze and align sub-sentence chunks. Concerning TMs covering Arabic, beside some commercial multilingual products in which Arabic is just one of the several languages provided, the most interesting example of an Arabic-English TM is Meedan (2009), an open access collection of several thousand paired text chunks extracted from Arabic newswires. Its structure is the simplest, providing just Arabic sentences paired with English translations, without any alignment of sub-sentences.

2.3 Combinatory-Categorial Grammar (CCG)

The choice of CCG as a grammatical paradigm to analyze and automatically translate idioms is based upon several grounds: (i) it is a perfectly formalized grammatical paradigm; (ii) some very performing implementations, such as OpenCCG (White, 2012; Bozşahin et al.; 2012), are available, with both parsing and generation capabilities; (iii) the lack of a theoretical status for phrase structure allows for highly unortodox structures to be represented, e.g. coordination among elliptical constructions (Steedman, 2000; Steedman and Baldridge, 2005), which fits well the complex nature of idioms requirements as far as phrase structure is involved; (iv) the combination of a very basic categorial apparatus with infinitely many complex categories and attributes allows for a smooth transition between open constructions, partially frozen collocations and more or less rigid idioms.

In the original, simplest version, A[jdukiewicz] B[ar-Hillel] Calculus (Bar-Hillel, 1953), a single "rule", functional application, is included: a complex category matches another element to its left or its right (according to the direction of the final slash) to form a larger category where the matched element is "erased" from the list of missing arguments. The function in the semantics of the complex category is applied to the semantics of the matched argument.

the po-liceman	departed
NP: **P**	S\NP: λx.**go**(x)
S: **go(P)**	

Example 1

the po-liceman	saw	the boy
NP: **P**	S\NP/NP: λxλy.**see**(y,x)	NP: **B**
	S\NP: λy.**see**(y,**B**)	
S: **see**(**P**,**B**)		

Example 2

AB Calculus is weakly equivalent to CF grammars (same generative power, possibly different analyses). This limitation does not allow the analysis of known phenomena that are slightly beyond strict context-freeness (e.g., cross formations in Dutch and Swiss German) and makes it difficult to handle unbounded dependencies. Since Curry & Feis (1958) "curried" operators (functional composition, type raising, crossed composition) have been introduced into the machinery of CG, which results in Combinatory Categorial Grammar (Steedman, 2000).

Related Work: Thanks to its very clear formal properties, CCG has been used for some very large implementations in parsing and generation. In particular, the CCGbank project (Hockenmaier and Steedman, 2005) translated the whole of Penn Treebank into a corpus of CCG derivations; the C&C CCG parser and supertagger, together with the Boxer computational semantics tool (Curran et al., 2007), have been explicitly designed for large-scale NLP tasks; OpenCCG, the OpenNLP CCG Library (Baldridge et al. 2007), implements a parser and a realizer with supertagging and hypertagging modules in the framework of multi-modal extensions to CCG (Baldridge and Kruijff, 2003). Several large grammars have been implemented in OpenCCG, including Moloko, a grammar oriented towards parsing and realization in human-robot interaction (Kruijff and Benjamin, 2012). However, with all their theoretical and empirical advantages CCG models have virtually never have been applied to the analysis of idioms nor to MT applications. The reason for this probably lies in a certain hesitation by linguists in the CCG framework to tackle language universals and in the idea that CCG semantic representation is best strictly coupled to its syntactic counterpart, which seems to make the treatment of wildly different structures that convey the same "meaning" in natural languages rather unlikely. As the proof-of-concept application presented in 3.2. shows, this is not necessarily the case.

3 The model and its implementation

The model we propose, given a list of IMWEs enriched by some semantic information, searches for them in collections of non sub-sententially aligned bilingual text (namely TMs), trying to pair each Arabic IMWE with the related translated chunk via the CCG representation module, that builds a syntactic-semantic representation of the matching IMWEs. The modular structure of the model will allow future developments, especially for the CCG component, which can be ideally extended in order to parse the entire TM and to get fully aligned bilingual versions.

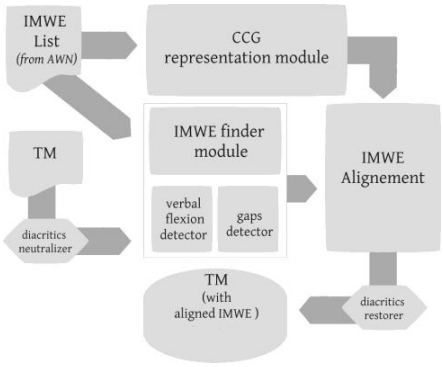

Figure 1: Model's process flow

3.1 Setting the MWE list and the pattern matching strategy

Since the aim of our model is not automatic extraction of MWEs, but rather testing alignment through a CCG interpretation, the IMWE list is a pre-existent input, but the condition is that every lexical entry must be previously associated with some semantic information (synonyms, English translations, ontological classification), usually available in networks such as Arabic WordNet (AWN: Black et al., 2006)). The main advantage in benefiting of data extracted from a lexical network is to have not only standard translations, but also all available MWEs in the target language. The example below shows a typical MWE entry used as an input:

intaqala 'ilā al-rafiq al-'a'lā (['die', 'decease', 'perish', 'go', 'exit', 'pass away', 'expire', 'pass', 'kick the bucket', 'cash in one's chips', 'buy the farm', 'conk', 'give-up the ghost', 'drop dead', 'pop off', 'choke', 'croak', 'snuff it'], ('die_v_1', 'Death'))

The length of the list is not very important, as the main task for this work is to have semantic data included in order to test the CCG module. Obviously

the model could benefit of other MWE sources, such as dictionaries, exhaustiive MWE repositories (see Hawwari et al., 2012), other network ontologies (e.g. Arabic VerbNet, FrameNet) or *ad hoc* lists built on web multilingual cross-referenced resources, such as Wikipedia.

At this stage of the model implementation we chose to focus on MWEs that contain at least a verb, in order to experiment more complex argument representations in CCG module. Almost all these MWEs share a low degree of compositionality and a certain morphosyntactic flexibility.

The other main input is obviously the TM, in which we would align the MWEs that match patterns in the list.

As it is known, the Arabic writing system includes a diacritization system for marking short vowels and consonant lengthening, but this system is rarely used in contemporary texts. However, since a partial diacritization is always possible even in contemporary writings, it can generate lots of false negatives if it is not taken into account. A small, independent module is therefore foreseen to neutralize full or partial vocalization in both MWEs list and TM processing and at the output stage to restore the original configuration.

Both inputs are then processed in a module that select the entries contained in MWE list as patterns to be matched in the TM. Since MWEs in the TM can have various degree of flexibility (namely verbal conjugation and a certain degree syntactic mobility of the constituents), two sub-modules has been conceived.

The first one accounts for morphological flexibility, but works in the lightest possible way, avoiding the need of new linguistic information. This is achieved by selecting in the verbal MWE pattern (always conjugated at past tense, third person masculine singular) those letters that are preserved in every conjugated form, i.e. the consonants (both belonging to the root and marking stems, e.g., *istaslama* > **s*t*s*l*m**, which is common to every conjugated form, such as *yastaslimu*, *istaslamna*, and so on). To deal with irregular verbs, the semi-consonants *w* and *y*, together with the *'alif* symbol are also ignored. In the next chapter it will be shown that this sort of brute-force method seems to provide better results than the employment of an external lemmatizer, namely the Buckwalter morphological analyzer (Buckwalter, 2002). Such tool appears to be instead more effective as a further strategy in refining results of the brute-force method, but this hypothesis was not yet tested with standard evaluation criteria.

The second sub-module simply allows to find MWE constituents in the target text even if they are intercalated with non MWE elements, by using gap detecting algorithms modeled on regular expression syntax.

Finally, the matching MWEs retrieved in the Arabic section of the TM are automatically tagged with the related source information contained in the original MWE list, in order to be processed by the CCG module.

3.2 Representation through CCG

As a proof of concept for the approach in representing syntax and semantics of idioms in the framework of a bilingual, bi-directional Arabic-English machine translation, two proof-of-concept (POC) grammars, one for each of the languages, were written in OpenCCG. Both grammars translate between surface forms and semantic representations and the other way round, being able to parse and generate from the same architecture. No direct language-to-language mechanism is included, and machine translation is rather a by-product of single-language parsing and generation facilities that share a common semantic representation.

The semantic representation avails itself of the dual representation level in OpenCCG: each non-purely functional word is grounded to a predicate and a class. The predicate is the main semantic value of a word and works as a key to parsing and, especially, generation. The class serves to match semantic restrictions on arguments: e.g., actors are animate.

In order to have a reasonably universal, or at least not excessively language-biased, semantic component, predicates are chosen among WordNet synsets and classes among SUMO concepts (Niles and Pease, 2001). These choices were induced by several reasons: on the one hand, WordNet (Miller, 1995) is perhaps the single most widespread lexical resource publicly available and a de facto standard in language technologies, alignments to it are available for many other resources —such as VerbNet, FrameNet, Wiktionary, SUMO and, most importantly, Arabic WordNet among many localized versions of the lexical database,— and it is a very practical choice for a universal semantic component; the unavoidable linguistic bias towards English will be overcome in further developments by treating WordNet as the main source for an International Language Index

(ILI: Vossen, 1998) together with other sources: this is what already happens in many localized WordNets, e.g. cultural-oriented concepts added in Arabic Wordnet are already assigned an ID distinct from the English WordNet.

On the other hand, the choice of SUMO concepts as a source for classes, though perhaps less straightforward, is reasonable as well; even if the roughly 3800 SUMO ontologies to which the 117k WordNet synsets are mapped are way too many for most reasonable linguistic tasks (VerbNet 3.2 uses only 36 selectional restrictions for 6031 verbs), the use of a larger ontology can be useful for more specialized lexical selection (e.g., the verbs in VerbNet class 38, animal sounds, all have the restriction [+animal] on the agent, but it is more reasonable also in linguistic terms to have a stricter restriction: for instance, only cats tend to meow) and —perhaps more importantly— the representation of the semantic component through ontologies with a rich axiomatization such as SUMO can be the input to further components, for instance a reasoner.

The POC grammar has a limited number of synsets, 5 nominal and 7 verbal ones, expressed by 18 English and 18 Arabic lexemes (including MWEs). The (rather large) subset of SUMO that encodes the corresponding classes, together with relevant WordNet synsets and English and Arabic lexemes, are shown in Figure 2 (arrows mark subclass relations, instanced classes have a light blue background, general classes for nouns and verbs are in salmon red and WordNet synsets are within boxes, with English and Arabic lexemes in italics).

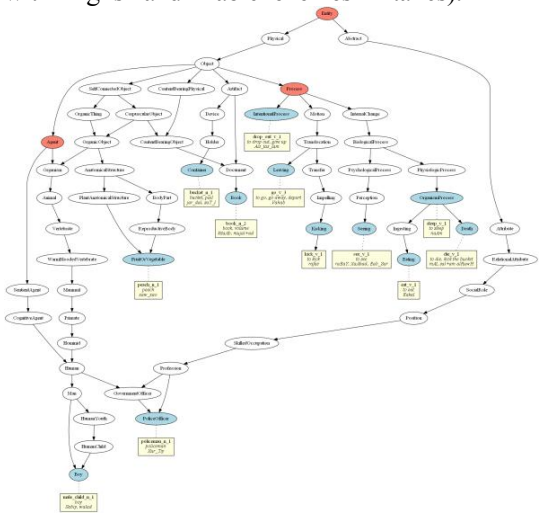

Figure 2: The network of SUMO classes, WordNet synsets and lexemes of the POC grammar

Despite its limitations, this POC addresses a number of potentially thorny issues in bilingual MT. First, the strongly lexical nature of CCG allows syntactic differences between English and Arabic to be abstracted away from semantic representation. E.g., the only relevant difference between Arabic and English intransitive verbs is the direction of the slash (basically, S/N in Arabic and S\N in English; we disregard here topic-initial sentences in Arabic, that are probably best analyzed as XVS structures according to the standard analysis in the Arabic grammatical tradition).

The key to extend the CCG approach to increasingly noncompositional lies in the more or less standard treatment of case-marking prepositions: if a verb requires a complement introduced by *to*, the latter does not contribute to the composition of the semantic representation; rather, it merely "checks" a syntactic feature that is needed for the derivation to continue.

In the same vein, the main significant element in an idiom is lexically assigned the semantic representation, while less significant elements are given a syntactic, checking function which is nevertheless necessary in order to let the derivation go on.

As an example, let us see how the system derives two idioms, one English and one Arabic, that Arabic WordNet considers equivalent to 'to die' in the meaning 'pass from physical life and lose all bodily attributes and functions necessary to sustain life', *kick the bucket* and *sal+am alruwH*[3], respectively (see entry example in 3.1). The English idiom admits of two reading, the idiomatic one and the less likely, but admissible, literal reading 'to give a kick to the pail'.

The POC grammar attributes the key role in the idiom to the verb *to kick* and uses the NP *the bucket* as a checking element. While debatable, this choice is not entirely arbitrary: on the one hand, it is *ceteris paribus* preferable to attribute the verb the key semantic role, since it already has the key role in the syntactic derivation; on the other hand, the shortened form *to kick* is attested in the meaning of the idiom, even if it is not recorded in WordNet (it is recorded in the English Wiktionary and in meaning I.b of $kick_{v1}$ in the OED).

The idiomatic and the literal derivations are shown in Figure 3 and Figure 4 respectively:

[3] The Arabic transcription is a 7-bit ASCII compliant version of the Buckwalter. We adopt a simplified morphology, without the final declension vowels that are usually omitted in everyday Modern Standard Arabic

Figure 3: derivation (idiomatic reading)

Figure 4: derivation (literal reading)

Two very distinct semantic representations are get by very similar syntactic derivations. The details of the semantic representations are in Figure 5 and Figure 6 respectively.

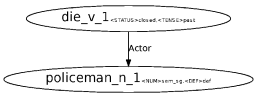

Figure 5: semantic representation (idiomatic reading)

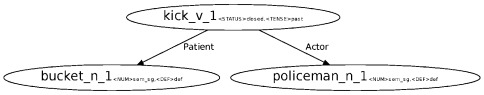

Figure 6: semantic representation (literal reading)

The Arabic version *sal+am alruwH* has basically the idiomatic reading only. The literal reading of 'to deliver the soul (to God)' is improbable enough, and close enough to the idiomatic reading, to have been excluded in our POC grammar (however, it might be included without altering the nature of the results). Here is the derivation for *sal+am AlXur_Tiy AlruwH* 'the policeman delivered his soul', i.e. 'died':

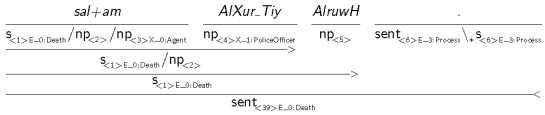

Figure 1: derivation of *sal+am AlXur_Tiy AlruwH*

The most striking feature of this derivation is that notwithstanding its radical syntactic dissimilarity from its English counterpart, it produces exactly the same semantic representation in Figure 6 above. This is the meaning of MT in this model: two or more sentences translate one another when they have the same semantic representation.

If we feed the English realizer with the representation in Figure 6 above we get the following sentences (in no particular order, unless we add scorer to the realizer, see White, 2012 for details):

the policeman died .
the policeman kicked the bucket .

If we feed the same representation to the Arabic realizer, we get

maAt AlXur_Tiy .
sal+am AlXur_Tiy AlruwH .

In both case, the first sentence is a literal, the second one an idiomatic, equivalent of *the policeman died* in the two languages.

On the other hand, if we feed the realizers with the representation in Figure 5, we get:

the policeman kicked the bucket .
the policeman kicked the pail .

for English, and:

rafas AlXur_Tiy Aljar_dal .
rafas AlXur_Tiy AlsaT_l .

for Arabic. In both cases, we have equivalents for the literal meaning of 'the policeman gave a kick to a (real) bucket', with different lexemes for 'bucket'.

This POC, notwithstanding its limitations, shows a series of interesting features: (1)identical meanings are captured despite of very different syntactic derivations, and different meanings are captured for the same input strings; (2) a language-independent representation of meaning is obtained, which can feed other components (reasoners etc.); (3) MT is a by-product of the parsing and realizing: translating in this model is not structurally different from paraphrasing (which is one of the main uses in current implementations of OpenCCG); (4) the system can be extended to other languages without the need to implement language-to-language grammar couples (the coupling is obtained through identity of semantic representations).

4 Testing model and results

4.1 Data and instruments

The source for the employed IMWE list is AWN (see also Rodrigo et al. 2008). Relatively small in size (it contains around 11,000 synsets), AWN utilizes the Suggested Upper Merged Ontology (SUMO) as a common interface to dialogue with previously developed wordnets.

We used two different TMs to test the model, one in Contemporary Arabic, the other in Classical Arabic. The first one is the Arabic-English Meedan

Translation Memory v.10 (Meedan 2009), which contains 59861 paired textual excerpts, mostly sentences, for around one million words. The source is declared to be newswires in Arabic.

The second one consists of our provisional version of a parallel Arabic-English corpus based on al-Bukhārī's collection of Hadiths. This corpus is still under development (and results are not still published) and at the present stage it only pairs the full *matn* (content) part with the correspondent English translation, without sentence segmentation. The number of paired *matn*s is 7305, with 382,700 words.

4.2 Testing and Results

At the beginning, all verbal MWEs have been extracted from the AWN verbal synsets, by searching for all entries containing at least one blank space surrounded by words. From the 666 resulting MWEs we omitted those with the pattern Verb + Preposition, as generally more compositional and less idiomatic. The resulting list was populated by 387 entries. To each entry its form without diacritics was then automatically associated. Both target TMs has been treated in the same way, by neutralizing any diacritization.

The MWE list fed the set of patterns to be searched in TMs (Meedan and Hadith corpus), with an interaction with related sub-modules to neutralize verbal conjugation and syntactic flexibility. The results of the MWE identification process are briefly shown in Table 2.

The automatic alignment of Arabic MWEs with correspondent English chunks was performed by using the drafted CCG module (results in Table 3.).

Results of both MWE retrieval in TMs and alignment through CCG have been submitted to standard evaluation practice. The two TMs were divided in training and testing sections, through division of each corpus in a training (85%) and testing (15%) part; the latter is currently still relatively small in consideration of the homogeneity of the corpus and the need to manually annotate the test sentences.

	MWE Retrieval		**CCG Alignement**	
	Meedan TM	Hadith TM	Meedan TM	Hadith TM
Error rate	20.57	15.35	8.07	10.9
Precision	85.24	88.29	95.68	93.23
Recall	94.19	96.36	96.25	95.87
F_1	89.49	92.15	95.96	94.53

Table 2 – Summary of results

Concerning MWE retrieval, a manual screening of the testing sample showed a consistent error rate (20.57% for Meedan TM and 15.35% for Hadith TM). However, considering the high number of false negatives (14.76% for Meedan TM and 11,71% for Hadith TM) compared to the small rate of false positives (5,81% for Meedan TM and 3,64% for Hadith TM) , the error rate seems to be mostly due to the relatively small size of the MWE list used as input (which can be easily extended) rather than to the effectiveness of the retrieval module and related sub-modules.

The results of the CCG processing and alignment of retrieved MWEs show instead that the model is highly efficient in pairing Arabic MWEs with related English translations.

5 Conclusions

The AraMWE project aims to bring together statistical analysis and extraction of MWEs in Arabic-English bilingual texts with MT and the building of a semantic representation of sentences containing idioms in the two languages. Although the project is still in its initial stage, preliminary results show the possibility to perform the retrieval stage of the task automatically and in order to feed a symbolic component whose general features have been successfully designed and tested.

Next stages in the project will involve the implementation of an Arabic-English bilingual grammar beyond the POC state in order to cope with a reasonable high percentage of sentences containing MWEs in aligned texts. The final aim of AraMWE is to build a hybrid system where a symbolic CCG-based core grammar is able to analyze, and to provide a semantic representation for, as large as possible an amount of relevant cases, by developing in parallel a statistical component which acts as a back-off mechanism for cases unrecognized by the symbolic component.

6 References

Al Khatib, Khalid, Amer Badarneh. 2010. Automatic Extraction of Arabic Multi-Word Terms. In Proceedings of IMCSIT-2010, 411-418.

Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In Proceedings of LREC-2010, Valletta, Malta.

Baldridge, Jason and Geert-Jan M. Kruijff. 2003. Multi-Modal Combinatory Categorial Grammar. EACL-03, 211-218.

Baldridge, Jason, Sudipta Chatterjee, Alexis Palmer, and Ben Wing. 2007. DotCCG and VisCCG: Wiki and Programming Paradigms for Improved Grammar Engineering with OpenCCG. In Proceedings of the Workshop on Grammar Engineering Across Frameworks. Stanford, CA.

Bar-Hillel, Yehoshua. 1953. A quasi-arithmetical notation for syntactic description. Language 29: 47–58.

Black, William, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease and Christiane Fellbaum. 2006. Introducing the Arabic WordNet Project, in Proceedings of the Third International WordNet Conference, Sojka, Choi, Fellbaum and Vossen eds.

Boulaknadel, Siham, Beatrice Daille, Driss Aboutajdine. 2009. A multi-word term extraction program for Arabic language. LREC 2008, 630–634

Bozşahin, Cem, Geert-Jan M. Kruij and Michael White. 2012. Specifying Grammars for OpenCCG: A Rough Guide.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, Philadelphia.

Cacciari, Cristina and Patrizia Tabossi, eds. 1993. Idioms: processing, structure, and interpretation. Lawrence Erlbaum Associates, Hillsdale, NJ.

Chuang, Thomas C., Jia-yan Jian , Yu-chia Chang , Jason S. Chang. 2005. Collocational Translation Memory Extraction Based on Statistical and Linguistic Information. *Computational Linguistics and Chinese Language Processing*, 10 (1), 329-346.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, Prague, Czech Republic. Association for Computational Linguistics, 41–48.

Curran, James R., Stephen Clark, and Johan Bos (2007). Linguistically Motivated Large-Scale NLP with C&C and Boxer. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), 33-36.

Curry, Haskell B. and Robert Feys. 1958. Combinatory Logic: Vol I. Amsterdam: North Holland.

Diab, Mona and Pravin Bhutada. 2009. Verb noun construction MWE token supervised classification. In *Workshop on Multiword Expressions (ACL-IJCNLP)*, 17–22.

Hawwari, Abdelati, Kfir Bar, Mona Diab. 2012. Building an Arabic Multiword Expressions Repository. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju, Republic of Korea, 12 July 2012. Association for Computational Linguistics, 24–29.

Hockenmaier, Julia and Mark Steedman. 2005. CCGbank: User's Manual. Technical Report MS-CIS-05-09, Department of Computer and Information Science, University of Pennsylvania, Philadelphia.

Kavka, Stanislav and Jerzy Zybert. 2004. Glimpses on the History of Idiomaticity Issues. In *SKAZE Journal of Theoretical Linguistics*, 1, 54-66.

Kruijff Geert-Jan M. and Trevor Benjamin. 2012. Documentation for the MOLOKO CCG grammar (v6). The DFKI Language Technology Lab.

Lagoudaki, Elina. 2006. Translation Memory systems: Enlightening users' perspective. Key finding of the TM Survey 2006 carried out during July and August 2006. Translation Memories Survey 2006. London, Imperial College.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, (11), 39-41.

Meedan. 2009. Meedan v.10. Arabic-English Translation Memory. Meedan, San Francisco, CA.

Moirón, Begoña Villada and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In the Workshop on Multiword Expressions in a Multilingual Context, EACL-06, Trento, Italy.

Niles, Ian and Adam Pease. 2001. Towards a Standard Upper Ontology. In Proceedings of FOIS-2001. Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19, 2001.

Pawley, Andrew. 1983. "Two puzzles for linguistic theory: nativelike selection and nativelike fluency." In: J. C. Richards and R.W. Schmidt (eds.), Language and Communication. London: Longman, 191-225.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. CICLing-2002: 1-15.

Simard, Michel. Translation Spotting for Translation Memories. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond* – Vol. 3, pp. 65-72.

Steedman, Mark and Jason Baldridge. 2005 Combinatory Categorial Grammar. In R. Borsley and K. Borjars, eds, Non-Transformational Syntax.

Steedman, Mark. 2000. The Syntactic Process, MIT Press. Boston, MA.

Vossen, Piek, ed. 1998. EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer, Dordrecht.

White, Michael. 2012. OpenCCG Realizer Manual.