A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields

M. Attia[1], M. Rashwan[1], A. Ragheb[1], M. Al-Badrashiny[1], H. Al-Basoumy[1]
[1] The Engineering Company for the Development of Computer Systems; RDI, Egypt
6th floor of Bank Missr building, 171st Al-Haram Av., 12111, Giza, Egypt
E-mails: {m_Atteya, Mohsen_Rashwan, Ragheb, Mohammed.Badrashiny, Basoumy}@RDI-eg.com

Abstract

Applications of statistical Arabic NLP in general, and text mining in specific, along with the tools underneath perform much better as the statistical processing operates on deeper language factorization(s) than on raw text. Lexical semantic factorization is very important in that aspect due to its feasibility, high level of abstraction, and the language independence of its output.

In the core of such a factorization lies an Arabic lexical semantic DB. While building this LR, we had to go beyond the conventional exclusive collection of words from dictionaries and thesauri that cannot alone produce a satisfactory coverage of this highly inflective and derivative language.

This paper is hence devoted to the design and implementation of an Arabic lexical semantics LR that enables the retrieval of the possible senses of any given Arabic word at a high coverage.

Instead of tying full Arabic words to their possible senses, our LR flexibly relates morphologically and PoS-tags constrained Arabic lexical compounds to a predefined limited set of semantic fields across which the standard semantic relations are defined. With the aid of the same large-scale Arabic morphological analyzer and PoS tagger in the runtime, the possible senses of virtually any given Arabic word are retrievable.

1. Introduction

This paper presents an Arabic lexical semantics LR that is composed of the following four logical components:

1- A compact basis set of predefined semantic fields; i.e. word senses.
2- Lexical semantics relational data base (RDB) where the Arabic lexical compounds from a given lexicon are one-to-many mapped to semantic fields both in the forward and backward directions.
3- A set of predefined standard semantic relations; e.g. antonymy, hyponymy, entailment ... etc.
4- An RDB connecting the semantic fields to one another via none, one, or multiple standard semantic relations.

In what follows; the need for this LR is first manifested in sec. 2. Next, the criteria that governed the design of the LR is manifested in sec. 3, hence the design itself is dissected in sec. 4 and the process of building the LR is explained in sec. 5.

Finally, sec. 6 compares this LR to the Arabic Word Net (AWN) which seems to be the most relevant one to ours.

2. Need for this LR

While the wide spectrum of text mining applications might perform patterns detection/comparison for many tasks by directly processing raw text, performance gets better and better as the mining is done on deeper and deeper linguistic analysis of this text given the same algorithms, training corpora, and computational power.

Mathematically, as we delve deeper in linguistic analysis (e.g. from morphological, to semantic …) resolving more and more complex relations, the raw text is factorized into more fundamental - and typically less numerous - atomic entities to be dealt with. This in turn reveals more concentrated statistical correlations and reduces the dimensionality of the problem, which both sharpen the effectiveness of the mining process. (Hearst, 1999; Jurafsky & Martin, 2000; Riloff & Jones, 1999; Schütze & Manning, 2000)

The importance of language factorization gets more and more magnified as the vocabulary and structure of the subject language gets richer. In fact, while Arabic is on the extreme of richness as per its vocabulary when regarded as full-form words, this language is also on the extreme of compactness of atomic building entities due to its systematic and rich derivative and inflective nature. (Attia, 2005; Attia, 2000; Dichy & Hassoun, 2005; Sulayman Fayyaadh, 1990) This positions language factorization not only as a performance boosting enhancement to Arabic text mining tasks, but also as a necessity for producing workable applications with useful output.

Among the most fundamental and feasible factorizations in this regard comes Arabic morphological analysis, Part-of-Speech (PoS) tagging, and lexical semantic analysis.

3. Design Criteria of the LR

In order to maximally satisfy the abovementioned need, our Arabic lexical semantic analyzer had to rely on an Arabic lexical semantics LR built according to the following criteria:

1- Originality of the source Arabic lexical semantic knowledge base. This means the LR, esp. its lexical side, should be designed in accordance with the intricate specifics of the Arabic language from the very beginning. This is an important missing feature in other lexical semantics LR’s like the Arabic Word Net (AWN). (Diab, 2004)

2- Widest coverage of possible Arabic lexical compounds, and semantic relations. Unless the highly derivative and inflective nature of Arabic is effectively handled, the runtime retrieval miss ratio of input words vs. the (inevitably limited) terms explicitly covered by the source of raw Arabic lexical semantics would be unacceptably high. So, this LR must go beyond the simplistic vocabulary-based model for maximally covering input Arabic terms and tying them to their possible senses/semantic fields.

3- Compactness of the resulting LR. Such an LR should never be huge in size not only in order to avoid prohibitive development, reviewing, and updating cost & time, but also to keep the LR development process from being excessively error-prone. So, this LR should be cleverly designed with a pacified growth of lexical/semantic relations versus the size of lexicon entries and semantic fields.

4- Independence and simplicity of the LR. Just like any professionally built LR; independence from the applications and from any LR development tools, as well as the simplicity of the LR format, are vital implicit aspects of this LR design.

5- Minimum implementation and updating cost. Less than 100 man-months within 2 calendar years had been allocated for building, refining, and verifying this LR. So, design decisions were always made in favour of the smaller, the clearer, the cleaner, and the faster choices. It was not always straightforward to satisfy this aspect together with the other ones of the criteria.

4. Design Description of the LR

To produce a sound Arabic lexical semantics LR complying with the abovementioned criteria, the implemented design relied on the following key concepts and choices:

4.1 Source of Raw Arabic Lexical Semantics

The published literature had been surveyed for sound semantic knowledge bases crafted originally for the Arabic language by specialized Arabic linguistics teams led by credible experts; (Dichy & Hassoun, 2005; Hanna Ghaleb, 2003; Ibraheem Al-Yazijy; La Rousse, 1988; Mahmoud I. Siny et al., 1993; Mukhtaar Umar et al., 2002; Rafael Nakhla; Sulayman Fayyaadh, 1990; Wagdy R. Ghaly, 1996).

Neatly based on the theory of semantic fields (Lehrer, 1974; Mukhtaar Umar, 1998), the Grand Thesaurus (Mukhtaar Umar et al., 2002) containing over 35,000 explicit Arabic lexical entries and relying on around 1,800 semantic fields has been elected to be our initial source of raw Arabic lexical semantics. Other sources are also used for the refinement and enrichment of the LR.

4.2 Arabic Lexical Compounds & Morpho-PoS Constraining

In order to avoid a prohibitively high runtime retrieval miss-ratio of input Arabic words versus the terms covered by the source(s) of raw Arabic lexical semantics,[1] Arabic *lexical compounds* and *morpho-PoS constraining* are introduced as two powerfully flexible concepts for taming the highly inflective and derivative nature of Arabic.

Instead of full-form words, the units of the lexical side in the lexical semantics DB of the LR are encoded as lexical compounds composed of the underlying morphemes that are flexible to be fully or partially matched against the morphemes composing the input words.

A morpheme code is explicitly mentioned *only if* its exact existence in the lexical compound is necessary to imply the semantic field(s) tied to this lexical compound. If the existence of *any* morpheme containing a certain PoS tag is only necessary to imply those semantic field(s), the code of this PoS tag with a negative sign is mentioned in place of that morpheme. A *don't-care* code (assigned -1000) in some place signifies that the morpheme at that place is semantically neutral.

Illustrative examples on morpho-PoS constrained lexical compounds are provided in tables 4 and 5 in sec. 5 below.

To realize such design concepts, RDI's Arabic morphological and PoS tagging factorization models are adopted in this LR. (Attia, 2005; Attia, 2000; Attia & Rashwan, 2004)

4.2.1. Arabic Morphological & PoS-Tagging Factorization Models from RDI

This Arabic morphological model assumes the canonical structure uniquely representing any given Arabic word $\boldsymbol{w}$ to be a quadruple of morphemes that $\boldsymbol{w} \rightarrow \boldsymbol{q} = (t{:}\ p, r, f, s)$ where p is prefix code, r is root code, f is pattern (or form) code, and s is suffix code. The type code t can signify words belonging to one of the following 4 classes: *Regular Derivative* ($\boldsymbol{w}_{rd}$), *Irregular Derivative* ($\boldsymbol{w}_{id}$), *Fixed* ($\boldsymbol{w}_f$), or *Arabized* ($\boldsymbol{w}_a$).

Prefixes & suffixes; P and S, the 4 classes applied on patterns; F_{rd}, F_{id}, F_f, and F_a, and only 3 classes applied on roots[2]; R_d, R_f, and R_a constitute together the 9 categories of morphemes in this model. The total number of morphemes of all these categories in this model is around 7,800. With such a limited set of morphemes, the dynamic coverage exceeds 99.8% measured on large Arabic text corpora excluding transliterated words.

While table 1 on the start of the next page shows this model in application on few representative sample Arabic words, the reader is kindly referred to (Attia, 2000) for the detailed documentation of this Arabic morphological factorization model and its underlying lexicon along with the dynamics of the involved morphological analysis/synthesis algorithms.

[1] The size of lexical entries in any such source has an order of magnitude of O($10^{4.5}$) while that of the generable Arabic lexical compounds via inflection and derivation is O(10^7).

[2] The roots are common among both the regular and irregular derivative Arabic words.

Sample word	*Word type*	*Prefix & prefix code*	*Root & root code*	*Pattern & pattern code*	*Suffix & suffix code*
فَمَا	*Fixed*	فَـ **2**	اَلَّذِي **87**	مَا **48**	ـ **0**
تَتَنَاوَله	*Regular Derivative*	تـ **86**	ن و ل **4077**	تَفَاعَلَ **176**	ـه **8**
اَلْكِتَابَات	*Regular Derivative*	الـ **9**	ك ت ب **3354**	فِعَال **684**	ـات **27**
اَلْعِلْمِيَّة	*Regular Derivative*	الـ **9**	ع ل م **2754**	فِعْل **842**	ـيَّة **28**
مِنْ	*Fixed*	ـ **0**	مِنْ **63**	مِنْ **118**	ـ **0**
مَوَاضِيع	*Regular Derivative*	ـ **0**	و ض ع **4339**	مَفَاعِيل **93**	ـ **0**
مُتَّخَذة	*Irregular Derivative*	ـ **0**	أ خ ذ **39**	مُتَّخَذ **13**	ـة **26**

Table 1: Exemplar Arabic morphological analyses.

On the other hand, our Arabic PoS-tagging model relies on a compact set of Arabic PoS tags containing only 62 tags covering all the possible atomic context-free syntactic features of Arabic words. While many of these Arabic PoS tags may have corresponding ones in other languages, few do not have such counterparts and may be specific to the Arabic language.

This PoS tags-set has been extracted after thoroughly scanning and decimating the morpho-syntactic features of the 7,800 morphemes in our morphologically factorized Arabic lexicon. Completeness, atomicity, and insurability of the scanned morpho-syntactic features were the criteria adhered to during that process.

Each morpheme in our Arabic factorized lexicon is labelled by a PoS tags-vector as exemplified by table 2 below:

Morpheme	*Type & Code*	*Arabic PoS tags vector label*
الـ	*P* 9	[Definitive] [ال التعريف]
سَيـ	*P* 125	[Future, Present, Active] [استقبال، مضارع، مبني للمعلوم]
مُفَاعِل	F_{rd} 482	[Noun, Subjective Noun] [اسم، اسم فاعل]
اسْتِفْعَال	F_{rd} 67	[Noun, Noun Infinitive] [اسم، مصدر]
مَلَائِك	F_{id} 29	[Noun, No SARF, Plural] [اسم، ممنوع من الصرف، جمع]
هُوَ	F_f 8	[Noun, Masculine, Single, Subjective Pronoun] [اسم، مذكر، مفرد، ضمير رفع]
ذُو	F_f 39	[Noun, Masculine, Single, Adjunct, MARFOU'] [اسم، مذكر، مفرد، مضاف، مرفوع]
ـات	*S* *27*	[Feminine, Plural] [مؤنث، جمع]
ـونَهُمْ	*S* *427*	[Present, MARFOU', Subjective Pronoun, Objective Pronoun] [مضارع، مرفوع، ضمير رفع، ضمير نصب]
ـيَّتَانِ	*S* *195*	[Relative Adjective, Feminine, Binary, Non Adjunct, MARFOU'] [نسب، مؤنث، مثنى، غير مضاف، مرفوع]

Table 2: PoS labels of sample Arabic morphemes.

Due to the atomicity of our Arabic PoS-tags as well as the compound nature of Arabic morphemes in general, the PoS labels of Arabic morphemes are represented by PoS tags-vectors.

While the Arabic PoS-tagging of stems is retrieved from the PoS label of the pattern only, not the root's, the PoS-tagging of the affixes is obtained from the PoS labels of the prefix and suffix. So, the Arabic PoS-tagging of a quadruple corresponding to a morphologically factorized input Arabic word is given by the concatenation of its PoS labels of the prefix, the pattern, and suffix respectively after eliminating any redundancy.

While table 3 below shows the Arabic PoS-tagging of few sample words, the reader is kindly referred to (Attia & Rashwan, 2004; Attia, 2005 [3]) for the detailed documentation of this Arabic PoS-tagging model along with its underlying PoS tags-set.

Sample word	*Arabic PoS tags vector*
فَمَا	[Conjunction, Noun, Relative Pronoun, Null Suffix] [عطف، اسم، اسم موصول، لا لاحقة]
تَتَنَاوَله	[Present, Active, Verb ,Objective Pronoun] [مضارع، مبني للمعلوم، فعل، ضمير نصب]
اَلْكِتَابَات	[Definitive, Noun, Plural, Feminine] [ال التعريف، اسم، جمع ، مؤنَّث]
اَلْعِلْمِيَّة	[Definitive, Noun, Relative Adjective, Feminine, Single] [ال التعريف، اسم، نسب، مؤنَّث، مفرد]
مِنْ	[Null Prefix, Preposition, Null Suffix] [لا سابقة، حرف، لا لاحقة]
مَوَاضِيع	[Null Prefix, Noun, No SARF, Plural, Null Suffix] [لا سابقة، اسم، ممنوع من الصرف، جمع ، لا لاحقة]
مُتَّخَذة	[Null Prefix, Noun, Objective Noun, Feminine, Single] [لا سابقة، اسم، اسم مفعول، مؤنَّث، مفرَد]

Table 3: PoS tags-vectors of sample Arabic words.

4.3 Packaging Format of the LR

All the components of this LR are formally structured as relational databases (RDB) which guarantees both its independence and simplicity.

4.4 Multi-Level Indirect Semantic Mapping

Instead of the infeasible direct semantic mapping of the whole Arabic vocabulary across itself with a size complexity of $O(V^2)$; V is the huge vocabulary size of Arabic, our LR is designed for the multi-level semantic mapping; $\boldsymbol{w_i \leftrightarrow LC_m \leftrightarrow SF_u \leftrightarrow SF_v \leftrightarrow LC_n \leftrightarrow w_j}$.

Input Arabic words $\boldsymbol{w_i}$ are analyzed into morpho-PoS constrained lexical compounds $\boldsymbol{LC_m}$ which are in turn mapped in the inverse direction of the lexical semantics RDB to semantic fields $\boldsymbol{SF_u}$.

The semantic fields are semantically interrelated through an $S \times S$ matrix per each defined semantic relation; where S is the size of the predefined basis set of semantic fields.[4] The third step of the mapping is hence possible.

Navigating our lexical semantics RDB in the forward direction can infer the possible $\boldsymbol{LC_n}$ that correspond to the semantic fields $\boldsymbol{SF_v}$ resulting from the previous step.

Morphological and PoS-tagging models help again at the last link in the chain of indirect semantic mapping across all the generable Arabic words.

[3] Esp. see chapter 4.

[4] This size has typically an order of magnitude of O($10^{3.5}$).

Given that $S<<V`<<V$; where $V`$ is the number of core lexical compounds mentioned explicitly in our LR, the size complexity of the indirect semantic mapping approach is then $O(S^2+S\cdot V`)=O(S\cdot V`)$ which is much more tractable than $O(V^2)$ of the direct semantic mapping.

5. The Building Process of the LR

The sources of raw Arabic lexical semantics knowledge base are usually organized so that the semantic fields/word senses are the primary keys that recall the terms belonging to them. Assuming such sources, the process of building our Arabic lexical semantics LR proceeds as follows:

1- After adding each distinct semantic field in the raw source to the basis set of semantic fields, the terms belonging to each field are linguistically reviewed to explicitly add/remove any missing/irrelevant terms under this semantic field.

2- Each of these Arabic terms is analyzed to obtain its morphological as well as PoS-tagging factorization, and is hence encoded as a morpho-PoS constrained lexical compound as previously explained in sec. 4.2.

3- This lexical semantic knowledge base obtained so far is then formally structured as an RDB with the semantic fields acting as the primary keys. This is called the *forward* Arabic lexical semantics RDB of which table 4 below shows a sample fragment.

Semantic Field	***Lexical Compound***	***Morphological–PoS Tagging Constraints of Lexical Compound***							***Meta Semantic Fields***	
		Q₁					***Q₂***	**…**		
		t	*p*	*r*	*f*	*s*				
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(التّأْلِيف), Authoring, 466	…	.	…	…	…	…	…	…	…	…
	أُطْرُوحَةٌ	1	-1000	2484	785	-48	×	×	▨	▨
	بَحْثٌ	1	-1000	211	817	-1000	×	×	▨	▨
	تَأْلِيفٌ	1	-1000	128	526	-1000	×	×	▨	▨
	رِسَالَةٌ	1	-1000	1565	684	-48	×	×	▨	▨
	مُؤَلَّفٌ	1	-1000	128	519	-1000	×	×	▨	▨
	سِجِـلٌّ	2	-1000	1893	208	-1000	×	×	▨	▨
	سِفْرٌ	1	-1000	1964	842	-1000	×	×	▨	▨
	كِتَابٌ	1	-1000	3354	684	-1000	×	×	▨	▨
	كِتَابَةٌ	1	-1000	3354	684	-48	×	×	▨	▨
	مَبْحَثٌ	1	-1000	211	792	-1000	×	×	▨	▨
	زَبُورٌ	1	-1000	1728	671	-1000	×	×	▨	▨
	مَخْطُوطٌ	1	-1000	1155	779	-1000	×	×	▨	▨
	مُصَنَّفٌ	1	-1000	2364	519	-1000	×	×	▨	▨
	…	.	…	…	…	…	…	…	…	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4: A fragment of the forward lexical semantic RDB.

The detailed documentation of the building process of this forward Arabic lexical semantics RDB, (written by Attia et al.) is freely downloadable at: http://www.rdi-eg.com/rdi/Downloads/Process_of_building_the_forward_Arabic_Lexical_Semantic_DB.pdf.

4- Using SQL operations, this forward RDB is automatically inverted so that the lexical compounds act as the primary keys. A sample fragment of this *inverse* lexical semantic RDB is shown by table 5 on the next page.

5- While building the inverse RDB, a special *back-off* row is inserted per each distinct root in the inverse RDB in order to further attenuate the runtime retrieval miss ratio of input words. The lexical compound of a back-off row mentions only the root morpheme explicitly, and all the other morphemes (prefix, pattern, and suffix) as *don't care*.
If an input word matches none of the explicitly registered derivatives of some root in the inverse RDB, the corresponding back-off row is resorted to. The recalled semantic fields of such a row are the union of the recalled semantic fields of all the registered derivatives of its root in the inverse RDB.

6- The basis set of semantic fields are interrelated via a matrix per each predefined standard semantic relation. So far, in addition to *relatedness* the following 20 semantic relations (Mukhtaar Umar et al., 2002) are defined in our Arabic lexical semantic LR:

1- *Antonymy*.
2- *Approximate Synonymy*.
3- *"Whole→Part"* relation.
4- *"Part→Whole"* relation.
5- *Hyponymy; "is-a-special-type-of"* relation.
6- Inverse of no. 5: *"is-a-general-type-of"* relation.
7- *Hyponymy; "is-a-member-of"* relation.
8- Inverse of no. 7: *"includes-several"* relation.
9- *Hyponymy; "is-originated-from"* relation.
10- Inverse of no. 9: *"is-the-origin-of"* relation.
11- *Hyponymy; "is-integrally-included-in"* relation.
12- Inverse of no. 11: *"includes-integrally"* relation.
13- *Causality: "is-a-cause-of"* relation.
14- Inverse of no. 13: *"due-to"* relation.
15- *Conditionality; "is-conditional-on"* relation.
16- Inverse of no. 15: *"is-a-condition-for"* relation.
17- *Temporal locality*: *"is-a-time-for"* relation.
18- Inverse of no. 17: *"occurs-during"* relation.
19- *Spatial locality*: *"is-a-place-of"* relation.
20- Inverse of no. 19: *"takes-place-in"* relation.

7- The totality of these matrices is then unified in one formal RDB compatible of the format of our LR.

It should be noted that the development team followed a *cross-checking* policy for ensuring the quality of this LR whose first edition had been completed in Oct. 2007 followed by a more refined one in mid. 2008.

Lexical compound												*Possible semantic fields*						
String	*Q₁*					*Q₂*					…	***SF₁***	***SF₂***	***SF₃***	***SF₄***	***SF₅***	***SF₆***	…
	t	*r*	*f*	*p*	*s*	*t*	*r*	*f*	*p*	*s*								
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
ك ت ب	1	3354	-1000	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	الاستثمار 1800	العُقود 450	المراسَلة 179	التأليف 466	الكِتابة 1678	الإلزام 590	…
تَكَاتَبَ	1	3354	176	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	العُقود	-	-	-	-	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
اِكْتَتَبَ	1	3354	249	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	الاستثمار	-	-	-	-	-	…
اِكْتَتَبَ فِي	1	3354	249	-1000	-1000	3	42	132	-1000	-1000	…	الاستثمار	-	-	-	-	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
اِكْتِتَاب	1	3354	280	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	الاستثمار	-	-	-	-	-	…
كَاتَبَ	1	3354	457	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	المراسَلة	-	-	-	-	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
مُكَاتَبَةٌ	1	3354	487	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	المراسَلة	-	-	-	-	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
كَتَبَ	1	3354	859	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	العُقود	المراسَلة	التأليف	الكِتابة	الإلزام	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
كِتَابٌ	1	3354	648	-1000	-1000	-1000	-1000	-1000	-1000	-1000	…	العُقود	المراسَلة	التأليف	الكِتابة	الإلزام	-	…
كِتَابَةٌ	1	3354	648	-1000	-48	-1000	-1000	-1000	-1000	-1000	…	المراسَلة	التأليف	الكِتابة	-	-	-	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 5: A sample fragment of the inverse Arabic lexical semantic RDB.

6. Comparison with AWN

Disseminated LR's relevant to ours, esp. Word Nets and Thesauri are surveyed. (Black et al., 2006), (Diab, 2004), (Ghonaimy, 2003), (Vossen, 2002), (Visual Thesaurus; http://www.VisualThesaurus.com)

Among those LR's; the beta release of the Arabic Word Net (AWN) www.GlobalWordNet.org/AWN/, www.LDC.UNPENN.edu, announced in Mar. 2007, has apparently been found closest and hence been thoroughly investigated and compared to our Arabic lexical semantics LR.

Feature	AWN	Our LR
Underlying theories	*Semantic Fields*, and *Componential Analysis of Semantic Fields*	*Semantic Fields*, and *Componential Analysis of Semantic Fields*
Format of LR	Hierarchical	RDB
Current no. of lexical entries	≈ 12,038	≈ 40,000
Current no. of semantic fields	5,861	1,824
Semantic relations defined	*Hyponymy* only	20 semantic relations (see sec. 5 above)
Auxiliary technologies	None	*Morphological* and *PoS-Tagging* factorization
Back-off upon mismatches	None	To the semantic fields of the root
Mapping to other languages	Many; esp. *English*	None

Table 6: AWN vs. our Arabic Lexical Semantics LR.

Interestingly, each of the two has shown superior/inferior complementary aspects to the other. While the AWN has a richer taxonomized set of semantic fields, and can also map to sister Word Nets in other languages (esp. English), our LR on the other hand has much richer semantic relations, much more explicit lexical entries, and much lower miss-ratio versus input words due to lexical compounds, morph-PoS constraining as well as the back-off.

A concise comparison between the two LR's is given in table 6 on the opposing column.

7. Conclusion

This paper has presented a large-scale Arabic lexical semantics LR with a wide coverage of the huge generable Arabic vocabulary. Based on the theory of semantic fields, the raw source content of this LR is primarily drawn from the best experts' works crafted originally for the Arabic language.

Packaged as an RDB, the primary key of this LR in its inverse format is a morphologically and PoS-tags constrained lexical compound that provokes in real time, using an Arabic morphological analyzer and PoS tagger, the semantic fields it may belong to. The relatively compact predefined set of semantic fields addresses the most common, if not all, the context-free word senses.

The standard semantic relations are labeled in matrices across that set of semantic fields which, indirectly along with the morphological & PoS-tags constraining, enables semantically relating virtually any possible couple of Arabic lexical compounds.

The mostly language independent issue of disambiguating the retrieved word senses has deliberately been located outside the scope of the presented LR and left to the applications layer that can benefit from numerous works reported in the rich published literature on that concern.

8. Acknowledgements

Building this Arabic semantic LR has taken off in late 2005 out of the need for a reliable Arabic lexical semantic analyzer to power the diverse applications of the 3-year “Arabic Text Mining” project funded by the “Data Mining & Computer Modelling” (DMCM) “Centre of Excellence” (CoE) which is the first centre within the national R&D CoE’s initiative launched by the Egyptian Ministry of Communications and Information Technology www.MCIT.gov.eg/Centers_Excellence.aspx

9. References

I- References in English:

Attia, M., 2005, *Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications*. PhD thesis, Dept. of Electronics and Electrical Communications, Cairo University. http://www.RDI-eg.com/RDI/Technologies/paper.htm.

Attia, M., 2000, *A Large-Scale Computational Processor of the Arabic Morphology, and Applications*, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University. http://www.RDI-eg.com/RDI/Technologies/paper.htm.

Attia, M., Rashwan, M., 2004, *A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words*, Proceedings of the Arabic Language Technologies and Resources Int’l Conference; NEMLAR, Cairo 2004. http://www.RDI-eg.com/RDI/Technologies/paper.htm.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Fellbaum, C., 2006, *Introducing the Arabic Word Net Project*; http://NLPweb.kaist.ac.kr/gwc/pdf2006/74.pdf.

Diab, M., 2004, *The Feasibility of Bootstrapping an Arabic Word Net Leveraging Parallel Corpora and an English Word Net*, Proceedings of the Arabic Language Technologies and Resources Int’l Conference; NEMLAR, Cairo 2004.

Dichy, J., Hassoun, M., 2005, *The DINAR.1 (DIctionnaire INformatisé de l’ARabe, version 1) Arabic Lexical Resource, an outline of contents and methodology*, The ELRA news letter, April-June 2005, Vol.10 n.2, France.

Ghonaimy, M.A., 2003, *A Tutorial Review on Word Nets*, Proceedings of the 4th Conference on Language Engineering; CLE’2003, the Egyptian Society of Language Engineering (ESLE).

Hearst, M., 1999, *Untangling Text Data Mining*, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), http://www.sims.Berkeley.edu/~hearst/papers/acl99/acl99-tdm.html.

Jurafsky, D., Martin, J.H., 2000, *Speech and Language Processing; an Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing*, Prentice Hall.

Lehrer, A., 1974, *Semantic Fields and Lexical Structures*, Amsterdam-London.

Riloff, E., Jones, R., 1999, *Learning Dictionaries for Information Extraction Using Multi-level Boot-strapping*, Proceedings of AAAI-99.

Schütze, H., Manning, C.D., 2000, *Foundations of Statistical Natural Language Processing*, the MIT Press.

Vossen, P., 2002, *Euro Word Net; General Document, Version 3, Final*, University of Amsterdam, http://www.hum.uva.nl/~ewn.

Yaseen, et al., 2006, *Building Annotated Written and Spoken Arabic LR’s in NEMLAR Project*, LREC2006 conference http://www.lrec-conf.org/lrec2006 , Genoa-Italy, May 2006.

II- References in Arabic:

[Abdul Kareem H. Gabal, 1997] "**في علم الدلالة**"، د.عبد الكريم حسن جبل، دار المعرفة الجامعية، الإسكندرية، 1997.

[Hanna Ghaleb, 2003] "**كنز اللغة العربية** "، د. حنا غالب، لبنان ناشرون، 2003م.

[Ibraheem Al-Yazijy] "**نجعة الرائد في المترادف والمتوارد** "، إبراهيم اليازجي، مكتبة لبنان، بيروت.

[Ibraheem Anees, 1952] "**دلالة الألفاظ**"، د. إبراهيم أنيس، مكتبة الأنجلو المصرية، 1952.

[La Rousse, 1988] "**المعجم العربي الأساسي** "، المنظمة العربية للتربية والثقافة والعلوم، لاروس، 1988م.

[Mahmoud I. Siny et al., 1993] "**المكنز العربي المعاصر**"، د. محمود إسماعيل صيني – وآخرون، مكتبة لبنان، بيروت، الطبعة الأولى، 1993م.

[Mukhtaar Umar, 1998] "**عِلْمُ الدِّلَالةِ**"، د. أحمد مختار عمر، عالَمُ الكُتُبِ، الطَّبْعَةُ الخَامِسةُ، 1998م.

[Mukhtaar Umar et al., 2002] "**المَكْنَزُ الكَبيرُ**"، د. أحمد مختار عمر – وآخَرُونَ، دارُ نَشْرِ "سُطُور" المملكة العربية السعودية، الطَّبْعَة الأُولَى، 2002م.

[Rafael Nakhla] "**المنجد في المترادفات والمتجانسات** "، الأب رفائيل نخلة اليسوعي، دار المشرق.

[Sulayman Fayyaadh, 1990] "**الحُـــقولُ الدِّلَاليَّةُ الصَّرْفيَّةُ لِلأَفْعالِ العَرَبيَّةِ**"، سُلَيْمان فَيَّاض، دارُ المَرِّيخِ بالرِّياضِ، 1990م.

[Wagdy R. Ghaly, 1996] " **معجم المترادفات العربية الأصغر** "، وجدي رزق غالي، مكتبة لبنان، بيروت، الطبعة الأولى، 1996م.