Mining the Web for the Induction of a Dialectical Arabic Lexicon

Rania Al-Sabbagh, Roxana Girju
Department of Linguistics, University of Illinois at Urbana Champaign,
USA
E-mail: alsabba1@illinois.edu, girju@illinois.edu

Abstract

This paper describes the first phase of building a lexicon of Egyptian Cairene Arabic (ECA) – one of the most widely understood dialects in the Arab World – and Modern Standard Arabic (MSA). Each ECA entry is mapped to its MSA synonym, Part-of-Speech (POS) tag and top-ranked contexts based on Web queries; and thus each entry is provided with basic syntactic and semantic information for a generic lexicon compatible with multiple NLP applications. Moreover, through their MSA synonyms, ECA entries acquire access to MSA available NLP tools and resources which are considerably available. Using an associationist approach based on the correlations between word co-occurrence patterns in both dialects, we change the direction of the acquisition process from parallel to circular to overcome a bottleneck of current research on Arabic dialects, namely the lack of parallel corpora, and to alleviate accuracy rates for using unrelated Web documents which are more frequently available. Manually evaluated for 1,000 word entries by two native speakers of the ECA-MSA varieties, the proposed approach achieves a promising F-measured performance rate of 70.9%. In discussion to the proposed algorithm, different semantic issues are highlighted for upcoming phases of the induction of a more comprehensive ECA-MSA lexicon.

1. Introduction

Arabic – the 6th official language of the United Nations (UN) – is a Semitic language, officially spoken in 22 countries in the Arab World with approximately 280 million native speakers and 250 million non-native speakers. It is divided into three major varieties: (1) Classical Arabic – the language variety of the Quran, (2) Modern Standard Arabic (MSA) – the language variety of most media, and (3) Colloquial Arabic or Arabic dialects, the native Arabic varieties. Most of research on Arabic has focused on MSA being the lingua franca and the official language variety in the Arab World; however, more recently much attention is also being given to dialectical Arabic which is not only a spoken variety now, but it is currently used as the written language variety of many websites, forums, blogs and some newspapers and books.

Recent research on Arabic dialects focuses on developing NLP tools – like, parsers (Chiang et al. 2005), POS taggers (Duh and Kirchhoff, 2005), diacritizers (Bakr et al, 2008) – rather than basic resources like lexicons. Therefore, it seems that most – if not all – existent dialectical Arabic lexicons are task-oriented as they include only the features that serve a particular NLP tool under development. Such features range from POS tags, root-and-pattern morphological templates or MSA synonyms.

Building NLP resources like lexicons is a computationally expensive process, especially for languages lacking enough corpora and basic language processing tools. In general, text collections available for Arabic dialects are sparse; even widely used dialects like Egyptian Cairene Arabic (ECA) still have small corpora in comparison to Modern Standard Arabic (MSA), for instance. As a result, many attempts to build NLP lexical resources, like lexicons, for Arabic dialects either use some manual assistance or restrict the lexicon to one lexical and/or syntactic feature avoiding broad coverage resources. Even when leveraging from available MSA resources, manual intervention and feature restriction remain the most dominating characteristics of the resources built for Arabic dialects.

In this paper, we try to induce a generic lexicon for ECA using a modified version of Rapp's (1999) algorithm for inducing word synonyms from unrelated documents based on correlations between word co-occurrence patterns. ECA entries are mapped to their MSA synonyms and thus they get access to the quite available NLP tools and resources for MSA. Moreover, each entry is tagged for the lexical features of gender and number which are acquired through the MSA synonym and each entry is given its top-ranked contexts acquired through Web queries.

This paper falls in seven parts. The first part briefly introduces ECA and its lexical inventory to clarify the type of ECA entries being included in our first phase. The second part reviews related work to dialectical Arabic lexicon induction in order to present the theoretical background of some of the techniques used in the proposed approach hereby. The third part highlights the lexicon induction approach and discusses in detail the circular collocation acquisition approach to overcome data sparseness in Arabic dialects. The algorithm is informally outlined in the fifth part of the paper. The evaluation methodology and metrics are spelled out in the sixth part. Finally, results are discussed and the contributions of the proposed approach are pointed out and future work is outlined.

2. Egyptian Cairene Arabic (ECA)

Egyptian Arabic is a continuum of four sub-dialects, geographically divided as:

- Coastal Egyptian of coastal towns like Alexandria,
- Lower Egyptian spoken in the Nile delta,
- Upper Egyptian spoken in Upper Egypt

- Egyptian Cairene Arabic (ECA) spoken mainly in the Egyptian capital, Cairo.

Among these, ECA is the most prominent given that it is understood all through Egypt and across most of the Arab World countries due to the predominance of the Egyptian media, basically movies, TV programs and series.

ECA vocabulary inventory can be divided into two major categories: ECA-exclusive vocabulary and ECA-MSA shared vocabulary. For the first phase of the ECA-MSA lexicon, this paper focuses solely on the dialect-exclusive vocabulary and leaves ECA-MSA shared vocabulary for further research where inter-dialect disambiguation techniques need to be developed as discusses in section 8.

ECA-exclusive vocabulary can be subdivided into three categories: acoustically-mitigated MSA vocabulary, foreign borrowings and dialectically-coined vocabulary. Since ECA is an everyday dialect which is generally used in informal situations, ECA speakers frequently mitigate MSA hard sounds, leading to the acoustically-mitigated MSA words that are used in spoken and sometimes written ECA. Table (1) shows the MSA hard sounds mitigated in ECA.

Arabic Sound		**Examples**	
MSA Hard Sound	**ECA Mitigated Sound**	**MSA Word**	**ECA Mitigation**
ذ */*/*	ز */z/*	ة */*mp/* (*protection*)	مة */zmp/* (*protection*)
ذ */*/*	د */d/*	كذب */k*b/* (*lying*)	كدب */kdb/* (*lying*)
ض */D/*	د */d/*	ضحكة */DHkp/* (*a laugh*)	دحكة */dHkp/* (*a laugh*)
ظ */Z/*	ض */D/*	ظلمة */Zlmp/* (*darkness*)	ضلمة */Dlmp/* (*darkness*)
ق */q/*	ا */A/*	قرنبيط */qrnbyT/* (*cauliflower*)	ارنبيط */ArnbyT/* (*cauliflower*)
ث */v/*	ت */t/*	ثمن */vmn/* (*price*)	تمن */tmn/* (*price*)
ث */v/*	س */s/*	ثانوي */vAnwy/* (*secondary*)	سانوي */sAnwy/* (*secondary*)
ء */'/*	Ø	افتراء */AftrA'/* (*injustice*)	افترا */AftrA/* (*injustice*)
ئ */}/*	ي */y/*	جائز */gA}z/* (*probable*)	جايز */gAyz/* (*probable*)

Table (1): Examples of MSA Hard Sounds Mitigated in Arabic

Foreign Borrowings – taken from different languages like English, French, Italian, Turkish and Spanish, form a considerable part of the ECA vocabulary exemplified in Table (2).

ECA Word	**Foreign Origin**	**MSA Synonym**
مرسي */mrsy/*	Merci (French)	شكراً */$kran/* (thanks)
أوضة */>wDp/*	Oda (Turkish)	حجرة */Hgrp/* (room)
دكتور */dktwr/*	Doctor (English)	طبيب */Tbyb/* (doctor)
جمبري */gmbry/*	Gamberetti (Italian)	روبيان */rwbyAn/* (shrimp)

Table (2): Examples of Foreign Borrowings in ECA

ECA Word Foreign Origin MSA Synonym

In addition to the acoustically-mitigated MSA vocabulary and foreign borrowings, there is dialectically-coined vocabulary that is originally coined as a part of the ongoing development of ECA. ECA-specific vocabulary includes both content and function words like: the noun حوسة */Hwsp/*[1] (*confusion*), the verb اتحنجل */AtHngl/* (*manipulate*), the adjective شعنون */$Enwn/* (*crazy*), the modifier شوية */$wyp/* (*some*), the interrogative pronoun إزاي */<zAy/* (*how*), the relative pronoun and complementizer اللي */Ally/* (*that*) among many others.

These three subcategories of ECA-specific words are the main target of the proposed algorithm for the first phase of our lexicon.

3. Related Work

Although NLP Research on Arabic dialects is still in its infancy, previous NLP work on Arabic dialects focused on building NLP tools rather resources and thus building task-oriented resources rather than generic ones with multiple applications.

Starting with an initial POS lexicon for ECA created by AraMorph (Buckwalter, 2002) – an MSA morphological analyzer – Duh and Kirchhoff (2005) developed a POS tagger for ECA. The importance of their attempt that is directly related to our approach hereby is that AraMorph (Buckwalter 2002) manages to provide possible POS tags for 62% of the ECA corpus which is a considerable recall rate, yet the accuracy rate was as low as 62.76%. This finding is specifically related to our approach hereby which shows with considerable common vocabulary between the two varieties, we can expect considerable common contexts as well that are to be used for lexical mapping, using cross-dialectical measures. The latter measures assume considerable commonality between Arabic dialects and the fact that the some dialects like LA have many resources.

[1] Buckwalter Transliteration Scheme

With the same idea in mind about cross-dialects commonalities, Rambow et al. (2005) used Rapp's (1999) algorithm for inducing lexical mappings from unrelated corpora based on the correlation of co-occurrence patterns of words in both corpora, using the city block distance as their similarity measure and log likelihood as their collocational association measure. No specific results were given about the performance of using Rapp's algorithm on Arabic dialects; however, Rambow et al. (2005) concluded that since their algorithm was much below the accuracy rate achieved by Rapp (1999) this algorithm may not be suited to the constraints on dialectical Arabic resources, given its small corpora and even small initial seed lexicons.

However, they had two major problems: seed lexicon and parallel acquisition of word co-occurrences. They initiated their algorithm with a manually compiled seed lexicon, which in addition to manual effort involved limits the output of the proposed algorithm to a specific path given the limited corpora of Arabic dialects. The problem of parallel acquisition of word co-occurrences is again sparse corpora and thus there algorithm is to achieve low results. Even with the improvements suggested by Rapp (2009a, 2009b) the performance remains the same and almost none of these improvements is applicable to Arabic dialects given that we don't have dictionaries and to reduce the dimensionality of the co-occurrence matrix using Singular Value Decomposition is also a trade-off between reducing noise, and reducing recall given the already small matrixes to be compiled from the small available corpora.

4. Lexicon Induction Approach

The lexicon to be built hereby is a generic lexicon which can be used for many NLP applications and tools. For each ECA entry, an MSA synonym is given and thus ECA is given access to multiple MSA tools and resources currently available. Moreover, each ECA entry is given within its context which is crucial to many NLP applications, including but not limited to Word Sense Disambiguation. Moreover, each ECA entry is tagged for Part-of-Speech (POS) and semantic features which also make the lexicon important for syntactic and semantic parsers[2].

The approach used for the lexicon induction is the same associationist approach used by Rapp (1999, 2009a, 2009b) among others where one word is defined in terms of its co-occurring words in local contexts. The underlying assumption of the associationist approach is that the co-occurrence patterns of words that are the same translation correlate. Yet, large corpora are required for precise correlation statistics.

Though using the Web as corpus is generally regarded as an appropriate approach to overcome data sparseness problems, using large corpora from diverse domains and genres, using the Web as corpus will partially handle data sparseness problem in terms of Arabic dialects. This is because Web content of Arabic dialects is still relatively small, in comparison to the Web content of MSA, for instance. In addition to the relatively small size of Web documents, the indexing techniques of Web documents on meta-search engines are usually problematic.

Results from highly visited Web documents are usually duplicated and since many – if not all – search engines limit their search results to 1,000 pages; many possible word co-occurrences are likely not to be found using regular Web query methodologies. As a result, we propose hereby a new technique for the acquisition of word co-occurrences using what we call Circular Acquisition instead of Parallel Acquisition.

In Parallel Acquisition (PA), word co-occurrences of each dialect/language variety are acquired separately and then they are matched later on using similarity measurements. This technique requires large corpora for both dialects/varieties and non-duplicated search results. Given the lack of these two conditions in Arabic Dialects Web content, we use instead Circular Acquisition (CA).

In CA, the acquisition of word co-occurrences of one dialect is conditioned by the word co-occurrences acquired for the other dialect. In other words, word co-occurrences in the first dialect (hereby ECA) are acquired first and then they are validated as possible co-occurrences for the words of the second dialect (hereby MSA). This approach is mainly intended to make sure that all word co-occurrences of the target dialect (i.e. ECA) are mapped to likely MSA synonyms.

A practical example on word co-occurrence found only through CA and not PA is found through the co-occurrences of the ECA adjective غلط /*glT*/ (*wrong*) and its MSA synonym خطأ /*xT>*/ (wrong). Using PA to acquire their co-occurrences, no similarities are found despite being absolute synonyms. For instance, search results for خطأ /*xT>*/ (*wrong*) include co-occurrences like: معلومات خطأ /*mElwmAt xT>*/ (*wrong information*) with 3,450 (Google Search Hits) GSHs and أفكار خطأ /*AfkAr xT>*/ (*wrong ideas*) with 197 GSHs; but neither معلومات /*mElwmAt*/ (*information*) nor أفكار /*AfkAr*/ (*ideas*) is found co-occurring with غلط /*glT*/ (*wrong*) in spite of being intuitively expected co-occurring words. This intuition is proved by submitting معلومات غلط /*mElwmAt glT*/ (*wrong information*) and أفكار غلط /*AfkAr glT*/ (*wrong ideas*) as search queries, yielding 7,610 and 5,710 GSHs, respectively. Therefore, CA is expected to overcome the some indexing limitations of search engines, given the relatively small Web content of Arabic dialects.

Another methodology that modifies Rapp's (1999) algorithm here is dispensing with the initial seed dictionary. In addition to limiting manual work, using such a dictionary will limit the possibilities of the acquired word co-occurrences to the semantic range of the words included in the seed dictionary. Instead, lists of ECA and MSA

[2] See appendix A for a sample of the compiled lexicon

words are randomly compiled and synonyms in both lists are being mapped together.

5. The Algorithm

The proposed algorithm relies on the immediate context of the ±1 word to define word co-occurrences using Web unrelated documents. Using immediate local context is motivated by previous observations (Dagan 1991) that immediate local contexts provide most of the lexical information about the target words, especially those necessary for word sense disambiguation. Moreover, using immediate contexts is expected to depict two syntactic relations between verbs and subjects, on one hand, and nouns and adjectives, on the other hand. In Arabic, the subject precedes the verb and the noun precedes the adjective and thus subjects and nouns can be contextual indicatives for the following verbs and adjectives and vice versa.

Pointwise Mutual Information (PMI) – defined in equation (1) below – is used to measure association between word co-occurrences instead of the log-likelihood ratio used in Rapp's (1999) original algorithm. Although Rapp (1999) used log-likelihood ratio, for being a better measure than the then-alternative Chi-square, PMI is expected to perform better given the sparseness of ECA data and the fact that PMI performs well for rare events. For similarity measures, the city-block metric is used to follow the same metric used by Rapp (1999) and its implementation on Arabic dialects (Rambow et al. 2005).

$$(1)\quad PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

Informally, the proposed algorithm can be described as follows:

```
1. Compile the immediate contexts of ECA words using a the search-engine queries
2. Measure the association between each ECA word and its compiled immediate contexts using PMI
3. IF collocation association score between the ECA word and its immediate context is > 0, THEN
    a. Search the web to validate ECA immediate contexts as possible contexts for the MSA words
    b. Measure the association between those immediate contexts and MSA words
    c. IF collocation association score between the ECA word and its immediate context is > 0, THEN
        i. This investigated immediate context is a valid context for MSA
4. Measure the similarity between the ECA and MSA vectors using the city-block metric defined as:
```

$$(2)\quad s = \sum_{i=1}^{n} |X_i - Y_i|$$

Adapt the semantic and POS tag of the MSA word – which is acquired using online MSA dictionaries – to their synonymous ECA words.

Using such an algorithm, 1,000 ECA words are mapped to their MSA synonyms, tagging them for the semantic features of gender and number and the appropriate POS.

6. Evaluation Matrics and Methodologies

Comparative evaluation with previously similar compiled dictionaries or with previous algorithms applied to ECA is not available since to the best of our knowledge this is the first study applied to ECA-MSA pair. Instead, two evaluation methogologies are being used: comparing against a baseline model and manual, human evaluation. Results of both methodologies are evaluated against three evaluation matrics: precision, recall and F-measure, respectively.

6.1. Baseline

Given that AraMorph (Buckwalter 2002) in Duh and Kirchhoff (2005) produced tags for 62% of their ECA corpora with a precision rate of 62.7%, it is used here as the baseline for the proposed algorithm. Particularly, AraMorph is used to evaluate two aspects of the proposed algorithm, namely POS tagging and the semantic features of gender and number given that AraMorph does not give ECA-MSA synonyms.

6.2. Human Evaluation

Manual evaluation is also used to evaluate POS tagging, semantic features labeling and synonym mappings between ECA and MSA. Two human raters are used with Kappa Coefficient as the measure for inter-rater agreement. It is formally defined as:

$$(3)\quad k = \frac{P(a) - P(e)}{1 - P(e)}$$

P(a) is the relative observed agreement among raters and *P(e)* is the hypothetical probability of chance agreement

6.3. Evaluation Matrics

Three evaluation matrics are used, namely recall, precision and F-measure which are defined as follows:

$$(4)\quad Recall = \frac{\#\ ECA\ entries\ mapped\ to\ MSA\ synonyms}{Total\ \#\ of\ ECA\ entries}$$

$$(5) Precision = \frac{\#\ ECA\ entries\ correctly\ mapped}{Total\ \#\ of\ ECA\ entris\ mapped\ to\ MSA\ synonyms}$$

$$(6)\ \mathbf{F-meaure} = \mathbf{2}\,.\,\frac{\mathbf{precision\,.\,recall}}{\mathbf{precision + recall}}$$

These evaluation matrics and methodologies are used on 300 ECA entries and their MSA synonyms.

7. Results and Discussion

Evaluated on 300 ECA entries and their MSA synonyms, the proposed algorithm hereby achieves the results summarized in table (3) below.

	Precision	**Recall**
AraMorph Baseline	0.424	0.32
Our Approach	0.72	0.7

Table(3) : Results of the Proposed Lexicon Induction Approach compared to Baseline Results generated by AraMorph

Recall errors are of two main types: 42% that yield no results on Web queries and 16% that yield collocational association score below our intuitively chosen threshold (i.e. the zero). This highlights the problem of data sparseness which remains partially unsolved despite using CA and the Web as corpus.

Another implication of sparseness of data is on precision. Some words are more frequently used only because they are more common and thus they are more likely to be found on Web documents that usually represent the most frequent forms of language usage.

Both recall and precision rates are also affected by the complex morpho-syntactic structures of ECA. Like MSA, ECA is a morphologically complex. For instance, a single ECA verb like اتعكش */AtEk$/* is a complete sentence, meaning 'he was arrested'. Ideally, it is to be mapped to the MSA تم اعتقاله */tm AEtqAlh/* (he was arrested). Handling such types of morpho-syntactically complex words in ECA is complicated because: first, it requires a morphological analyzer and a parser, none of which is currently available for ECA; second such MSA phrases are not found in machine-readable dictionaries and thus they need to be extracted from generic corpora though it is almost impossible to build a database of all possible MSA phrases.

Another reason for low precision is the very long dependencies in ECA where the subject of a Verb Phrase (VP) can be as far as –20 words and thus all surrounding –1 words are not relevant co-occurring words. There is a trade-off between widening the search space of word co-occurrences and the noise introduced to the algorithm input.

8. Conclusion and Future Work

In this paper, we presented an approach to build a lexicon of ECA and their MSA synonyms, their contextual synonyms. In particular, relying on a novel metric, the circular collocation acquisition, our model acquires the MSA synonym, POS tags and the semantic features of number and gender for 1,000 ECA entries with a promising F-measured performance rate of 70.9%.

The hereby ECA lexicon can be extended in two ways. First, Word Sense Disambiguation (WSD) techniques are to taken into consideration to deal with both inter-dialectical and intra-dialectical ambiguities. Inter-dialectical ambiguity is related to shared words between ECA and MSA, especially MSA words that acquire a new meaning when used in ECA or change their meanings. Intra-dialectical ambiguity is a common semantic problem where the same ECA word has more one meaning and sometimes more than one POS tag. WSD at both levels is made feasible by giving contextual synonyms rather than absolute synonyms for each of the hereby ECA entry; these entries can provide a training set.

Second, Multi-Word Expressions (MWE) are to be added to the induced lexicon. These expressions can be divided into three categories. The first includes MWE that literally consist of all MSA words, yet their non-literal meaning is dialectical as in ينزل من نظري */ynzl mn nZry/* (literal: *get out of my eyesight*; non-literal dialectical: *despise*). The second category includes expressions which consist of a mixture of MSA and ECA words like كرسي في الكلوب */krsy fy Alklwb/* (literal: *a chair in the lamp*; non-literal: *cause a problem*). Finally, there are MWEs that consist totally of ECA words like اخد في وشه *yAxd fy w$h/* (literal: *to take in his face*; non-literal: *to leave*).

With these two improvements into consideration and with ECA parsers and morphological tools being currently developed by other research groups, the proposed algorithm is expected to achieve higher performance rates and to build a more comprehensive ECA lexicon.

7. References

Bakr, H., Shaalan, K. and Zeidan, I. (2008). A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacriticized Arabic. *Proceedings of INFO2008*, Cairo, Egypt, 2008

Buckwalter, T. (2002). Arabic Morphological Analyzer (AraMorph). Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0

Chiang, D., Diab, M., Habash, N., Rambow, O. and Shareef, S. (2005). Parsing Arabic Dialects. Final Report, 2005 JHU Summer Workshop, 2005

Dagan, I. (1991). Lexical Disambiguation: Sources of Information and their Statistical Realization. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991

Duh, K. and Kirchhoff, K. (2005). POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, June 2005.

Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., Nichols, C. and Shareef, S. (2005). Parsing Arabic Dialects. Technical Report, the Johns-Hopkins University, 2005 Summer Research Workshop

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, MD, USA, 1999

Rapp, R. (2009). Automatic Dictionary Expansion Using Non-parallel Corpora. In *Advances in Data Analysis, Data Handling and Business Intelligence* (pp 317-325). Springer Berlin Heidelberg: USA.

Rapp, R. (2009). The Automatic Generation of Thesauri of Related Words for English, French, German, and Russian, *International Journal of Speech Technology,* 11, 3-4, 147-156

Appendix A:

This appendix gives a snapshot of some entries of our ECA-MSA lexicon, where ENTRY is the ECA target word; POS for the part-of-speech of the ECA word, SEM for the semantic features with +M meaning MALE and –P means SINGLUAR, MSA syn. in the MSA synonym(s), and CNTXT is the context of the target ECA entry.

ECA-MSA Lexicon Snapshot	**English Translation of the ECA-MSA Contents**
ECA ENTRY: ابتدي POS: V SEM: +M –P MSA Syn.:ابتدأ؛ بدأ CNTXT: عمري_ابتدى الوقت_ابتدى الاسلام_ابتدى المقطع_ابتدى الهبل_ابتدى	ECA ENTRY: start POS: V SEM: +M –P MSA Syn.: start; begin CNTXT: My age_strats, time_starts, Islam_strats, The scene_starts Madness_starts
ECA ENTRY: ابهات POS: N SEM: +M +P MSA Syn.: آباء CNTXT: ابهات_وامهات ابهات_الاطفال ابهات_الجيل ابهات_البنات	ECA ENTRY: fathers POS: N SMF: +M +P MSA Syn.: fathers CNTXT: Fathers_and mothers; Fathers_children Fathers_generation Fathers_daughters
ECA ENTRY: ابيح POS: ADJ SEM: +M –P MSA Syn.: فاسد؛ ماجن؛ منحل؛ CNTXT: كليب_ابيح فيلم_ابيح كلام_ابيح موضوع_ابيح	ECA ENTRY: immoral POS: ADJ SMF: +M –P MSA Syn.: immoral, spoilt, abusive CNTXT: abusive_clip immoral_movie, Immoral_speech Immoral_topic
ECA ENTRY: ابيضاني POS: ADJ SEM: +M –P MSA Syn.: ابيض اشقر CNTXT: تركي_ابيضاني طويل_ابيضاني لبناني_ابيضاني وسيم_ابيضاني	ECA ENTRY: fair POS: ADJ SEM: +M –P MSA Syn.: fair; blond CNTXT: fair_Turkish, fair_tall, fair_Lebanese, fair_handsome