Persian Idioms
Collection and Identification in Texts

Vahid Ostovar

Unitex workshop

2014

Introduction

Persian idioms with the structure class *C1* (Maurice Groos, 1996)

Persian :	وحید	تاب	خورد
Transliteration:	Vahid	tab	khord-past
Translation:	Vahid	swing	eat
Gloss:	Vahid had fun		

وحید, *Vahid is* the subject *(N0)*
تاب, *tab (swing)* is the direct object *(C1)*
خورد, *khord (eats)* is the verb *(V)*

Project aim

- The main goal of this paper was to prepare a database of Persian idioms that could be used in the computational processing of that language.

Objectives

- Selection of web sources for the collection of idioms
- Building a database of Persian idioms
- Building FST tools for corpus exploring
- Corpus parsing and idioms candidate extraction
- Evaluation

Methodology

- **Selection of web sources**

  - Queries on the web

  - Using Google and AVG * browsers

  - From January 2013 to May 2013

  - Systematically survey of results, the first top ten pages per query

* AVG: http://www.avg.com/eu-en/secure-search

These are the most relevant queries that have been used.

Q1)
[فرهنگ اصطلاحات عامیانه زبان فارسی]
Farhange estelahate amiane zabane Farsi
“Persian language expressions dictionary”
Q2)
[اصطلاحات عامیانه و رایج به زبان فارسی]
Estelahate amiane va rayej be zabane Fars
“common expressions in Persian language”
Q3)
[فرهنگ اصطلاحات زبان فارسی]
Farhange estelahate zabane Farsi
“expressions dictionary of Persian language”

Several hits were discarded as not relevant, namely:

- Sites about English idioms and other languages in Farsi.
- Sites that did not contain enough idiomatic expressions and were really poor to be considered as a source.
- Many of them just give some general information about Persian idiomatic expressions and advertise dictionaries of idioms that will be publish in the future.
- Websites that contains old Persian expressions.
- Sites appearing more than once in each query.

We retained 7 websites as a valid source and we collected 131, $C1$ type idioms from them and the other idioms that is 233 different idioms, collected from two dictionaries of Persian idioms.

Building a database of Persian idioms (class C1)

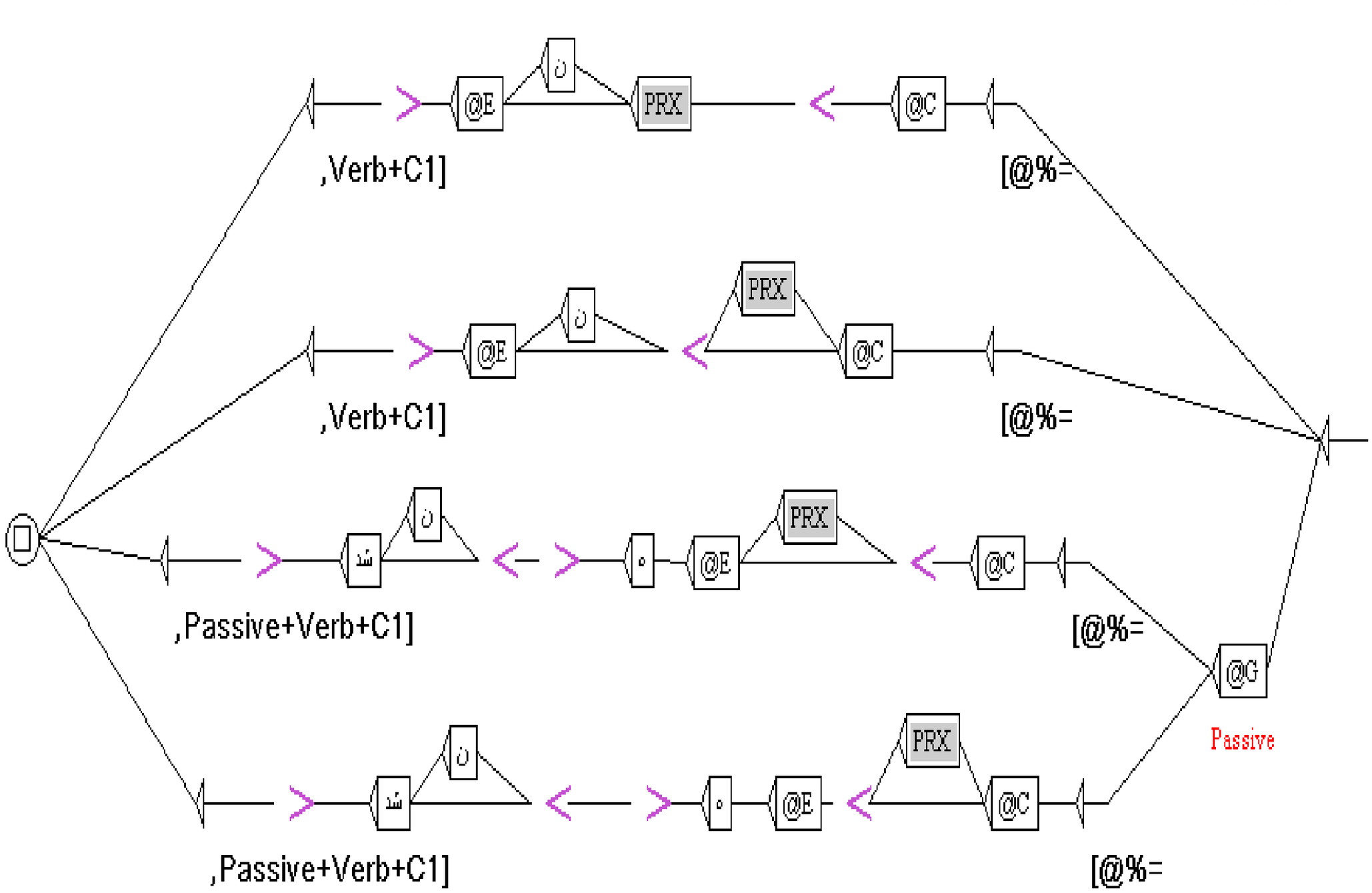
N0=:Hum	N0=:Nhum	C1	Det	Verb	Modify	Passive	Literal	Persian idioms	Translite ration	Gloss	Rough trans-lation	Exact mean-ing	Reference
+	-	آب	<E>	آوردن	<E>	-	+	آب آوردن	ab avardan	water bring	bringing water	being sick	http://www.scict.ir
+	-	آب	<E>	دادن	<E>	-	+	آب پس دادن	ab pas dadan	water give back	giving back the water	being generous -being unreliable	http://www.scict.ir
+	+	آب	<E>	رفتن	<E>	-	+	آب رفتن	ab raftan	water go	going water	became short- became low	http://www.scict.ir
+	-	آتش	<E>	کردن	<E>	-	+	آتش کردن	atash kardan	fire do	making fire	shooting	http://www.scict.ir

Building FST tools for corpus exploration

Intersecting the tabular data with finite-state transducer (**FST**), using the Unitex* (3.0) corpus processing tool.

Unitex*:http://www-igm.univ-mlv.fr/~unitex/index.php?page=0

Reference graph

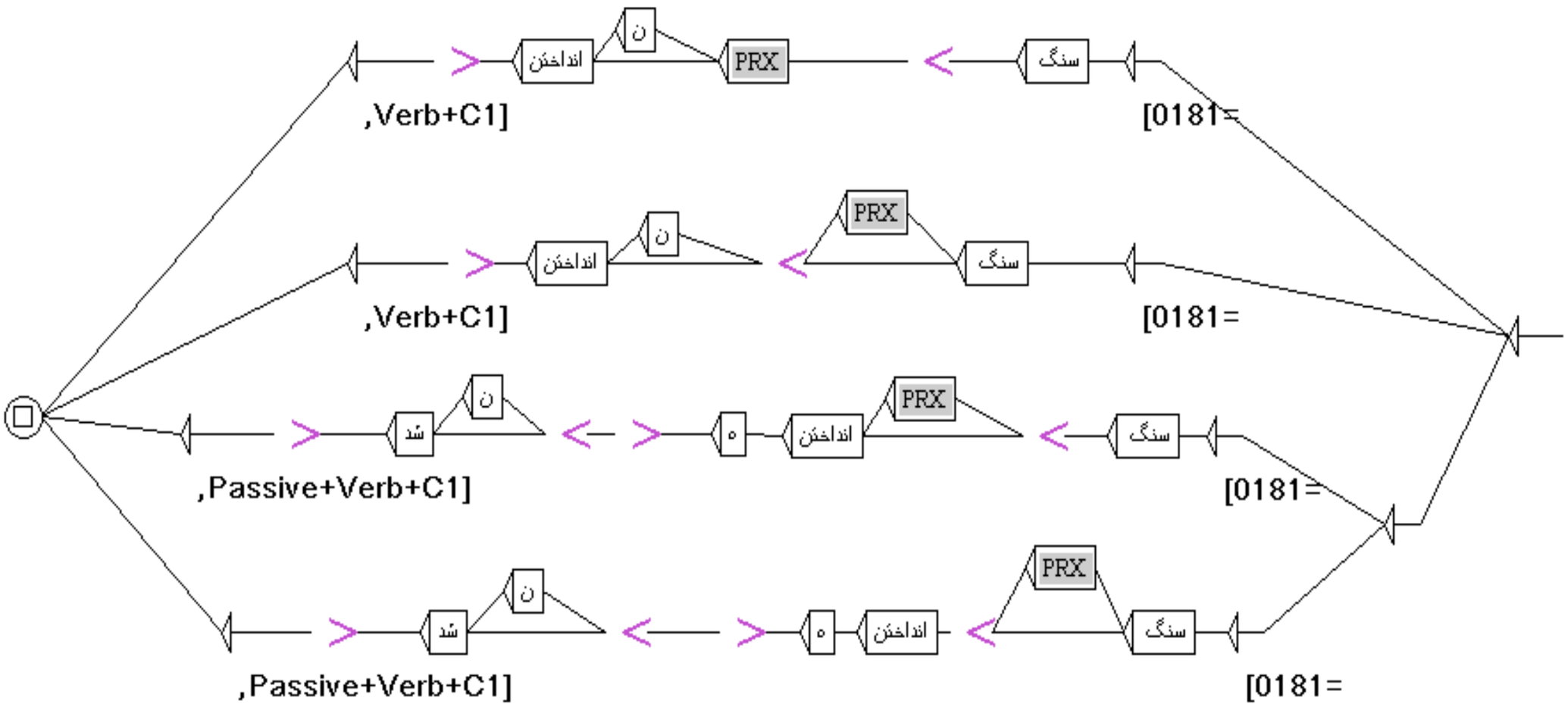


Intersecting the reference graph with the database

C1res.grf

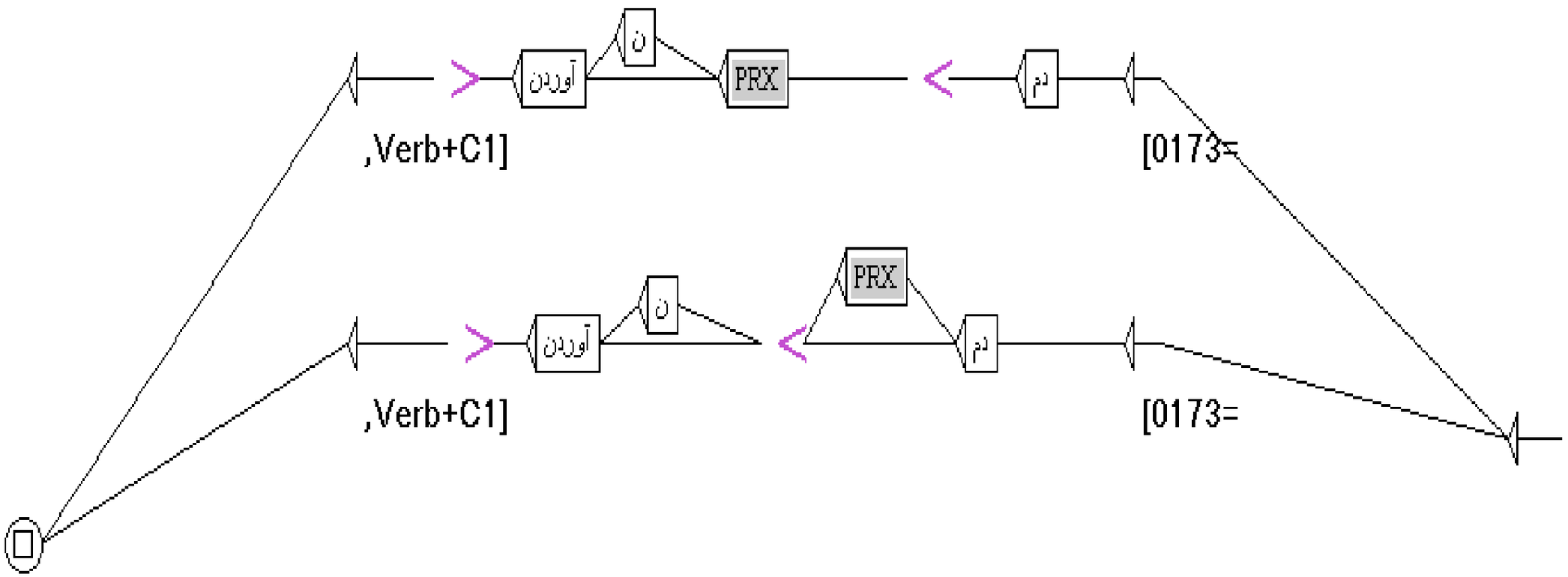
Persian : وحید سنگ انداخت
Transliteration: *Vahid sang andakht*
Translation: *Vahid is throwing stone*
Gloss: *Vahid is a trouble maker*

C1res.0181.grf



Persian: وحید دم در آورد
Transliteration: *Vahid dom daravard*
Translation: *Vahid evolves tail*
Gloss: *Vahid becomes shameless*

C1res.0173.grf



Concordance from TEP corpus

م . 100701 اما مايل نيستم كه با سرو گوش [Verb+C1,آب دادن=0002] زيادي براي همه مسئله درست كنم .
ابن به شما علامت ميدم . 462879 ابنا با [Verb+C1,آتش گرفتن=0005] دود غليظ سياهي از خودشون ميدن
شما چي فكر ميكنيد كه اين اتفاق به وسيل [Verb+C1,آتش گرفتن=0005] يك صفحه رخ داده بشه . 353933 م
219165 من دوستام را ديدم كه جلوي چشمام [Verb+C1,آتش گرفتن=0005] 219166 . [Verb+C1] كجا تو قايق . 219167
د علفهاي زير فيروو تي كه تو مزرعه بود ن [Verb+C1,آتش گرفتن=0005] 127736 . [Verb+C1] اون هم سال 1935 بود .
نه عزيزم تو هنوز هم نابسته اي . 322361 [Verb+C1,ياد كردن=0013] فقط يك امتياز اضافيه . 322362 ت
بس كن هرى سازمان كمابى را كه عمه شون را [Verb+C1,ياد كردن=0013] نميفرسته آزكابان . 24428 از طرف
. 63289 گمش نكني . 63290 من منظورشو از [Verb+C1,ياد كردن=0013] مارج نفهميدم من فقط . 63291 كنت
ار دوست داري اي . 99404 اينو يك بسر خوب [Verb+C1,يار آوردن=0014] تا هميشه فروتنانه سفارش بده خي
ريدم . 205391 ولي ببين چه افتضاحي به [Verb+C1,يار آوردن=0014] . 205392 جي از واگن ها دزد
8 نه ببخشيد مامان من يك عقيده محكم براي [Verb+C1,پس ندادن=0024] دارم . 148609 اگر نميتوني بري ت
نگ هي حمله كنيم . 543588 جانگ بوگ آماده [Verb+C1,تاب آوردن=0025] خسارات بزرگه . 543589 اين شرما
درت شنوايي و نه قدرت رواني . 87287 براي [Verb+C1,تاب آوردن=0025] برابر قدرت با عظمت صداي واقعي
46407 كمك رسي كمك . 46408 اميل شروع به [Verb+C1,تاب خوردن=0026] با لامپ . 46409 اميل به طرف من
اهزنهايي كه خودشون را به جايبرنس موكانگ [Verb+C1,جا زدن=0031] 424958 . [Verb+C1] همه جا هستند . 424959 بل
اغش به بدر من . 233216 ميدونستم كه براي [Verb+C1,جا زدن=0031] قطعات باهوشيم . 233217 اين كابربو
اريس . 16022 دنيال من بيان لطفا . 16023 [Verb+C1,حفظ كردن=0032] اين دو تا هرم . 16024 خيلي نادر

Corpus parsing and idioms candidate extraction

The corpus TEP**:** Tehran English-Persian Parallel Corpus

- from Natural Language and Text Processing Laboratory of Tehran University
- 4 million tokens on each side
- Sentence Aligned

Processing with Unitex 3.0

556.234 sentence delimiters
15.166.987 (64.492 diff) tokens
4.485.147 (64.365) simple forms
3.239.250 (10) digits

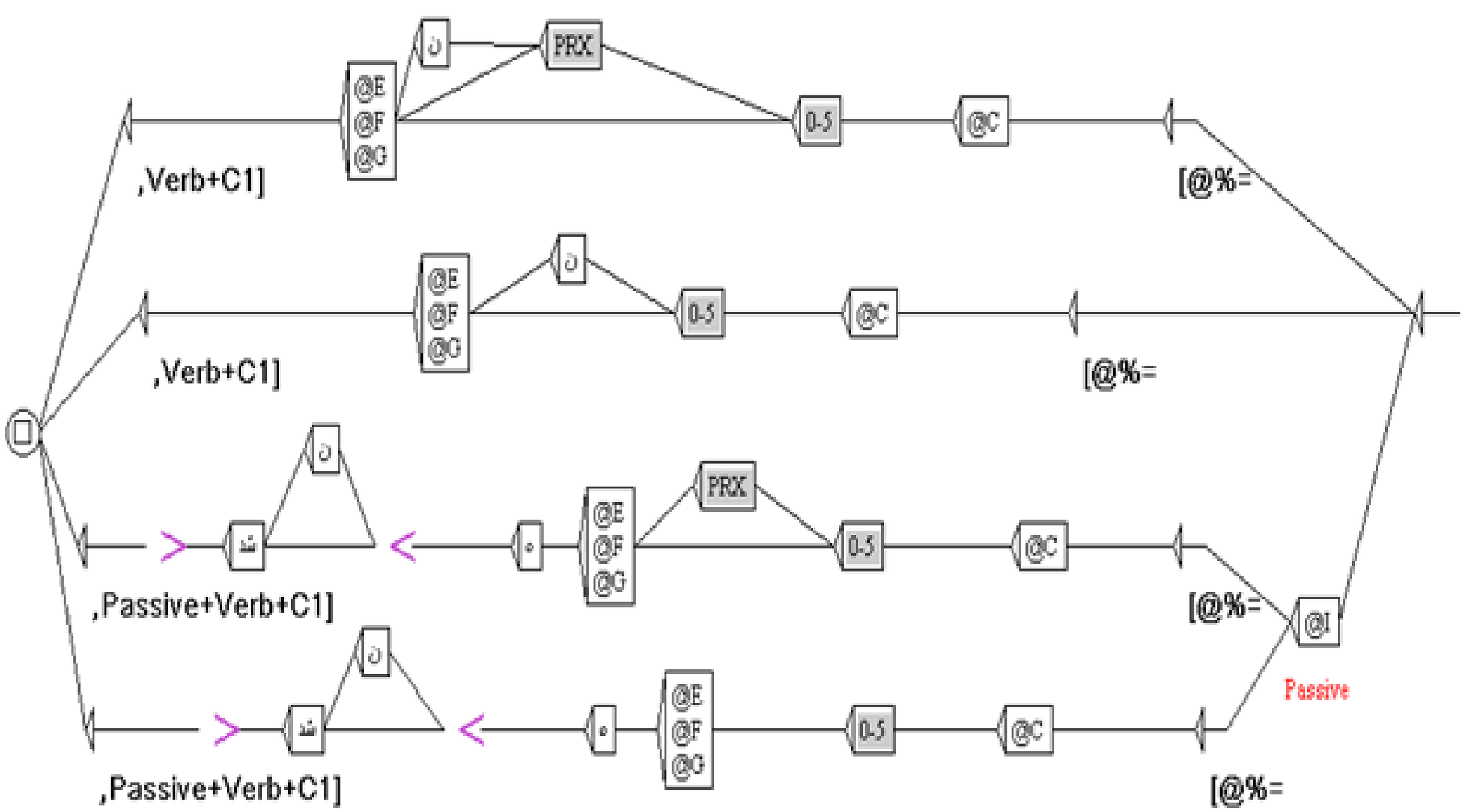
Evaluation

As the purpose of this project is to build a lexical database, the evaluation consist in three tasks:

1) Estimating the scope of the database
2) Determining the precision of the task of identifying the idioms, using the Unitex tools
3) Association measures

Estimation of idioms database scope

To estimate the scope of the database idioms, we produced this dummy FST to extract all the instances containing the verb and the lexical constant C1 from the text.



Database with the past and present stem of each idiom verb (the dummy FST is built according to this table)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
NO=:Hum	NO=:Nhum	C1	Det	Verb	past tense	present ten	Modify	Passive	Literal	Persian idioms	Translitera tion	Gloss	Rough translation	Exact meaning	Reference
-	+	برق	دع>	افتادن	افتاد	افت	دع>	-	-	برق افتادن	bargh oftad	electricity fall	falling electricity	being too shiny and clean	http://www.scict.ir
+	-	افلاس	دع>	افتادن	افتاد	افت	دع>	-	-	افلاس افتادن	eflas oftad	Bankruptcy fall	falling bankruptcy	Getting caught in the poverty	http://www.sootak.ir
+	-	پس	دع>	افتادن	افتاد	افت	دع>	-	-	پس افتادن	pas oftad	back fall	falling back	fainting	Persian proverbs and idiomatic expressions book
+	-	تک	دع>	افتادن	افتاد	افت	دع>	-	-	تک افتادن	tak oftad	monad drop	droping monad	being alone	Persian slang dictionar

Concordance of dummy FST

م . 100701 اما مايل نيستم كه با سرو گوش [0002=آب دادن,Verb+C1] زيادي براي همه مسئله درست كنم .
انواده اي كه حكومت ازش ميگه . 7801 اونا [0002=آب و نايش دادن,Verb+C1] اونا ضيطوني نخونش ميدن .
تيش درست كنيم . 95581 مي خواستم با بسرم [0004=آتش درست كردن,Verb+C1] را تمرين كنيم . 95582 باقه
هيم كرد . 311825 آماده شو براي انفجار و [0004=آتش روشن كردن,Verb+C1] همون جوري كه گفتم . 311826
ابن به شما علامت ميدم . 462879 اينا با [0005=آتش گرفتن,Verb+C1] دود غليظ سياهي از خودشون ميدن
شما چي فكر ميكنيد كه اين اتفاق به وسيل [0005=آتش گرفتن,Verb+C1] يک صفحه رخ داده بقه . 353933 م
219165 من دوستام را ديدم كه جلوي چشمام [0005=آتش گرفتن,Verb+C1] . 219166 كجا تو قايق . 219167
د علفهاي زير ضيروو تي كه تو مزرعه بود ن [0005=آتش گرفتن,Verb+C1] . 127736 اون هم سال 1935 بود .
ا مغزت بايد بيدار بقه . 87911 اونا من و [0011=ساب ساكت را جز جمع كردن,Verb+C1] . 87912 كدوم كير
نه عزيزم تو هنوز هم قابسته اي . 322361 [0013=ساد كردن,Verb+C1] فقط يک امتياز اضافيه . 322362 ت
بس كن هري سازمان كمابي را كه عمه شون را [0013=ساد كردن,Verb+C1] نميفرسته آزكابان . 24428 از طرف
. 63289 گمش نكني . 63290 من منظورشو از [0013=ساد كردن,Verb+C1] مارج نفهميدم من فقط . 63291 كنت
ر مترقبه جادويي فورا اعزام شدند . 63215 [0013=ساد اون را خالي كردن,Verb+C1] و حافظه اش را اصلاح
نفرم كه بيام . 82599 خوبين بر شما سلامت [0013=ساد دعا كردن,Verb+C1] بهتر است . 82600 امروز ميخو
ار دوست داري اي . 99404 اينو يک بسر خوب [0014=سار آوردن,Verb+C1] تا هميشه فرونشانه سفارش بده خي
ريدم . 205391 ولي ببين چه افتضاحي به [0014=سار آوردن,Verb+C1] . 205392 جي از واگن ها دزد
دتاتون با ورود به اين محاكمه . 8950 چند [0014=سار اسبش را در اين محاكمه آوردن,Verb+C1] . 8951 د
نگ هي حمله كنيم . 543588 جانگ يوگ آماده [0025=تاب آوردن,Verb+C1] خسارات بزرگه . 543589 اين فرما
درت شنوايي و نه قدرت روائي . 87287 براي [0025=تاب آوردن,Verb+C1] برابر قدرت با عظمت صداي واقعي

The results of the reference FSTs for idioms with the verbs کردن, *kardan (to do)* and زدن, *zadan (to hit)*, with and without insertions.

Experiment	Number of entries (types)	Number of matches	Different idioms (types)
Entire database with insertion	364	1,754	115
Entire database without insertion	364	584	68
کردن, kardan (to do) with insertion	57	510	36
کردن, kardan (to do) without insertion	57	107	23
زدن, zadan (to hit) with insertion	47	84	12
زدن, zadan (to hit) without insertion	47	60	7

Concordances result of each dummy FSTs

	کردن, kardan (to do)			زدن, zadan (to hit)		
	TP	FP	S	TP	FP	S
FST w/ Ins	34	81	395	49	18	17
FST w/o Ins	53	54	0	42	18	0

Determine precision of identification by using the Unitex tools

	کردن, kardan (to do)	زدن, zadan (to hit)
FST w/ Ins	0.29	0.73
FST w/o Ins	0.50	0.70

Association measures

For the idioms that were captured by the FST using the entire lexicon-grammar of *kardan (to do) and zadan, (to hit)*, we calculated two association measures; t-test and chi-square. (Manning and Schütze**,** 2005)

Idioms with the verb زدن, *zadan (to hit).* T-test and Chi-square

Idiom	Example (transliteration) translation/gloss	T-test	Chi-square	t > p	χ^2 > p
[0002]	جا زدن (ja *zadan*) fearing to do something	2.949175	30.63058	1	1
[0004]	گند زدن (gand *zadan*) sabotage	5.590053	2651.799	1	1
[0016]	قاط زدن (ghat *zadan*) becoming angry	1.402532	238.6465	0	1
[0023]	پرسه زدن (parse *zadan*) walking without a goal	0.896753	7.804938	0	1
[0026]	نیش زدن (nish *zadan*) Squibbing	1.970058	259.7837	0	1
[0035]	قاپ زدن (ghap *zadan*) sudden stealing	1.405453	319.5216	0	1
[0045]	جوش زدن (josh *zadan*) be concerned	1.288642	18.74106	0	1

Idioms with the verb کردن, *kardan (to do)*"T-test and Chi-square

Idiom	Example (transliteration) translation/gloss	T-test	Chi-square	t > p	χ^2 > p
[0001]	آتش کردن (atash *kardan*) shooting	-9.36316	8.575242	0	1
[0003]	باد در کردن (bad *darkardan*) farting	2.629322	21.82092	1	1
[0005]	تابلو کردن (tablo *kardan*) revealing something	0.641734	1.164957	0	0
[0007]	تُرش کردن (toesh *kardan*) becoming angry	1.78106	29.36749	0	1
[0008]	جفت کردن (joft *kardan*) fearing	-0.96383	0.479417	0	0
[0010]	کف کردن (kaf *kardan*) getting excited	-0.43306	0.132633	0	0
[0012]	باد کردن (bad *kardan*) getting pregnant	2.629322	21.82092	1	1
[0015]	قاطی کردن (ghati *kardan*) becoming angry	2.951347	544.3165	1	1
[0016]	کلید کردن (kilid *kardan*) being bête noire	0.734618	2.060893	0	0
[0017]	تَخم کردن (tokhm *kardan*) greatly fear	-1.5742	0.975672	0	0
[0018]	ردیف کردن (radif *kardan*) making everything alright	1.311004	23.86749	0	1
[0021]	یف کردن (pof *kardan*) feeling self-respect	2.357399	149.8065	0	1
[0022]	تَرش کردن (torsh *kardan*) huffing	1.78106	29.36749	0	1
[0027]	پنچر کردن (panchar *kardan*) being too tired	1.292239	19.62157	0	1
[0036]	اب باز کردن (ab baz*kardan*) happening something wrong	-5.20311	9.252444	0	1
[0038]	افاقه کردن (efaghe *kardan*) having good effect	0.986731	74.36311	0	1
[0040]	بچه درست کردن (bache dorost*kardan*) reproduction	-52.302	52.0676	0	1
[0041]	درز کردن (darz *kardan*) make known something private	0.814233	3.616851	0	0

Comparison of T-test and Chi-square

	T & χ^2 (1 1)	~T & χ^2 (0 1)	T & ~ χ^2 (1 0)	~(T & χ^2) (0 0)
کردن (kardan) to do	4	11	_	8
زدن (zadan) to hit	2	5	_	_

Conclusion

- Results indicate that depending on the verb, namely, if the verb has a more grammatical status or is a full verb, it may be necessary to adjust insertion windows between the key elements of the idiom.
- Precision varies depending on the verb.
- Association measures are valid strategies to capture these types of idioms; however, their precision depends on the idioms' verb, as the presence of support verbs may vary the results.

Future work

- Enhance the syntactic complexity of the reference graphs in order to capture other syntactic patterns such as *permutation*
- Plug the existing graphs with the Persian dictionary of inflected forms
- Extend the collection of idioms by using other sources, augmenting the lexical coverage of the lexicon-grammar of the idioms.
- Extend the collection of idioms to other types of Persian idioms


Thank
you