گزارشی از ساخت نخستین پیکره چند زبانه برای زبان فارسی

بهرنگ قاسمی‌زاده*، سعید رحیمی، مرتضی سالاریان، علی بهاری سلیم

چکیده

این مقاله، اولین پیکره فارسی موازی با تعداد زیادی از زبان های اروپایی را معرفی می نماید. این مقاله، اولین قدم ها را برای ساخت منابع اساسی جهت پردازش زبان فارسی معرفی می نماید. این مرحله از کار شامل معرفی ویژگی های صرفی-نحوی زبان فارسی و رمزنگاری آنها بر پایه مدل EAGLES/MULTEXT و منابع خاص MULTEXT-East می باشد. این مقاله پس از معرفی مختصر زبان فارسی، با تاکید بر شیوه نگارش و ویژگی های صرفی-نحوی آن، به ارائه یک دسته بندی جدید برای مقوله های گفتاری فارسی پرداخته و رسم الخطی را جهت ارائه فارسی در محیط دیجیتال ارائه می نماید. پس از آن آماری از پیکره ساخته شده ارائه می شود. کار صورت گرفته منحصرا توسط افراد داوطلب و بدون هیچگونه پشتیبانی مالی و یا معنوی از سازمان خاص صورت گرفته است.

واژه‌های کلیدی

پیکره، زبانشناسی رایانه ای، ویژگی های صرفی-نحوی، فارسی، حاشیه نویسی پیکره ها، پردازش زبان طبیعی.

Report of Constructing the First Parallel, Multilingual Corpus of Farsi

Behrang Qasemizadeh, Saeed Rahimi, Morteza Salarian, Ali Bahari Salim
Text and Speech Technology Corporation

Abstract

In this paper, we have introduced the first parallel corpus of Farsi with more than 10 other European languages. This paper describes primary steps toward preparing Basic Language Resources Kit (BLARK) for Farsi. Since now, we have proposed morphosyntactic features of Farsi based on EAGLE/MULTEXT models and specific resources of MULTEXT-East. The paper introduces Farsi Language, with emphasize on its writing system and morphosyntactic features, then a new Part-of-Speech categorization and orthography for Farsi in digital environments is proposed. Finally, the corpus and related static will be described. This work has been done in an informal way and with a volunteer based team.

Keywords

Natural Language Processing, Morphosyntactic Features, Orthography, Corpus, Computational Linguistics, Corpus Annotation, Parallel Corpora, Farsi

* سرپرست تیم پدیدآورندگان پیکره ، کارشناس ارشد هوش مصنوعی، شرکت فن‌آوری متن و گفتار، qasemizadeh@comp.iust.ac.ir

۱- مقدمه

امروزه با افزایش اهمیت فن آوری ارتباطات و اطلاعات، نیاز به فن‌آوری‌های زبان و گفتار نیز افزایش یافته است. برای اینکه افراد بتوانند از زبان مادری خود بر روی رایانه ها استفاده نمایند، نیاز به یک مجموعه تدارکات اولیه (همانند واژگان، پیکره و مجموعه ای از ابزارها) است. تلاشهای زیادی جهت مهیا نمودن بسته منابع زبانی پایه[۱] برای پردازش زبانها، به خصوص زبان هایی که از لحاظ تجاری کمتر مورد توجه اند صورت گرفته است. حاصل این تلاش ها، فراهم آمدن پیکره‌ها، ابزارهای تحلیل زبان همانند تحلیلگر صرفی و ... است. علاوه بر این، این منابع به راحتی قابل دسترسی و برای مقاصد آموزش و تحقیق کاملا رایگان است. متاسفانه برای زبان فارسی تا به حال کارهای مختصری در این زمینه صورت گرفته است. [۱][۲] علاوه بر آنکه کارهای صورت گرفته نیز مطابق با استاندارد و یا چارچوب کاری خاص صورت نگرفته است و اکثر آنها جهت استفاده محققین و دانشجویان در دسترس نمی باشد.

این مقاله به اولین قدم های صورت گرفته جهت فراهم آوردن منابع اولیه جهت پردازش زبان فارسی اشاره دارد. این شامل معرفی رسم الخط (قراردادی جهت ارائه پیکره) برای فارسی در محیط دیجیتال و رمزنگاری آن، معرفی مقوله های گفتاری، و معرفی ویژگی‌ها صرفی-نحوی فارسی منطبق بر مدل های ارائه شده توسط EAGLE/MULTEXT و استاندارد ارائه شده در MULTEXT-East است. [۳]

پروژه MULTEXT-East گسترشی از پروژه EU-MULTEXT بود که به گسترش منابع زبانی برای شش زبان شامل بلغاری، چک، استونی، لهستانی، رومانیایی، اسلونی و زبان انگلیسی به عنوان زبان مرکزی[۲] در پروژه پرداخت. نتیجه اصلی حاصل از این پروژه یک پیکره چند زبانه حاشیه نویسی شده، و مجموعه ای از ابزارها برای این هفت زبان است. این پروژه برای دیگر زبان های اروپایی نیز در حال گسترش و پیاده سازی است. کارآمدترین قسمت از این پروژه شامل منابع تشریح ویژگی های صرفی-نحوی(MSD)[۳] است که از سه لایه تشکیل شده است. این سه لایه به ترتیب انتزاع به شرح ذیل است[۵]:

- MSD 1984: پیکره ۱۹۸۴ (کتاب ۱۹۸۴ اثر جرج اورِل) که با ویژگی های صرفی-نحوی حاشیه نویسی شده است. در این پیکره هر یک از واژه ها با ویژگی های صرفی-نحوی رفع ابهام شده با توجه محل ظهور آنها، حاشیه نویسی شده اند.
- MSD Lexicons: واژگانی از واژه های صرف شده که در پیکره ظاهر شده اند به همراه لما و ویژگی های صرفی-نحوی آنها.
- MSD Specs: شرح ویژگی های صرفی-نحوی. این شرح مشخص می نماید که چه MSD هایی برای یک زبان معتبر است و معنای آن چیست برای مثال Ncms به معنی PoS اسم، نوع اسم عام، جنس خنثی، و عدد برار مفرد است.

بدین ترتیب MULTEXT-East چارچوبی قابل فهم برای گسترش پیکره ها فراهم می آورد. علاوه بر این منابع مختلفی مبتنی بر این چارچوب وجود دارد. به عنوان مثال پیکره ۱۹۸۴ برای چندین زبان مختلف وجود دارد. نکته جالب توجه اینکه ترجمه رمان۱۹۸۴ به فارسی وجود دارد و از سوی دیگر با توجه به ماهیت زبان فارسی به عنوان یک زبان هندی-اروپایی امکان اضافه کردن آن به این چارچوب امکان پذیر است. بنا بر آنچه که گفته شد و با توجه به نیاز های موجود برای پردازش رایانه ای فارسی اولین قدم ها جهت اضافه نمودن فارسی به این چارچوب برداشته شد. در کار گذشته نویسندگان این مقاله [۳] فاکتورهای مهم در این زمینه، شامل رسم الخط فارسی و نحوه رمزنگاری آن در محیط الکترونیکی، زبان فارسی و ویژگیهای صرفی و نحوی آن، و در نهایت تعامل میان این فاکتورها و چگونگی تاثیر آنها بر فرایند پردازش زبان به تفصیل بحث شده است.

در این مقاله، ادامه فعالیت های صورت گرفته در این جهت، یعنی ساخت پیکره فارسی و استفاده از تئوری های ارائه شده در [۳] در یک کاربرد واقعی گزارش شده است. با توجه به اهمیت موضوع در بخش ۲ ابتدا خلاصه ای از کارهای صورت گرفته ارائه می شود. نسخه جدید از آخرین تغییرات اعمال شده بر ویژگی های صرفی-نحوی و جداول مرتبط با آن مطالب بخش ۳ را تشکیل می دهند. بخش ۴ به ارائه پیکره ساخته شد و آمار مرتبط با آن می پردازد. بخش ۵ از مقاله به نتیجه گیری و شرح کارهای آتی اختصاص داده شده است.

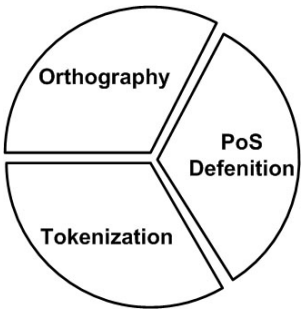
۲- زبان فارسی در چارچوب MULTEXT-East

همانطورکه گفته شد، در کار قبلی، زبان فارسی در چارچوب MULTEXT-East گنجانده و مطابق با آن ویژگی های صرفی-نحوی زبان فارسی ارائه شده است. در پاره ای موارد نیز با توجه به برخی ویژگی های خاص زبان فارسی، گروهی از ویژگی ها به چارچوب ذکر شده اضافه شده است. اضافه کردن فارسی به این چارچوب با چالش‌های عمده ای روبرو بود مهمترین چالش ها از این میان شامل موارد ذیل است:

- مشکلات مرتبط با رمزنگاری و ارائه فارسی در محیط الکترونیکی، و همزمان با آن رسم الخط فارسی در محیط الکترونیکی
- تعریف ویژگی های صرفی-نحوی مطابق با واقعیات زبان فارسی و در نظر گرفتن کاربرد تعاریف ارائه شده در زبانشناسی رایانه ای و تحلیل رایانه ای زبان فارسی
- تعامل و اثر متقابل تعاریف ارائه شده در بند ۱ و ۲، و همزمان با آن تاثیر بر فرآیندهای تحلیل همانند توکن بندی متون الکترونیکی فارسی و دیگر تحلیل های بعدی. (شکل ۱)

برای حل مشکل اول یعنی رسم الخط و رمزنگاری متون الکترونیکی، پس از مطالعه و مشورت با اساتید امر، تصمیم بر آن شد تا روش ارائه متون الکترونیکی فارسی ترکیبی از استاندارد ارائه شده توسط موسسه ملی استاندارد برای رمزنگاری کاراکترهای فارسی در

Unicode [۶] و همزمان با آن رسم الخط رسمی فارسی [۷] ارائه شده توسط فرهنگستان ادب و زبان فارسی برای سیستم مبتنی بر کاغذ (paper-based) باشد. در حقیقت روش ارائه شده برای رسم‌الخط الکترونیکی فارسی تطابق [۷] با [۶] است. مطابق روش ارائه شده، کاراکتر کنترلی ZWNJ به جای نیم‌فاصله در متون ظاهر می‌شود و مطابق آن کلیه پیش‌وندها و پسوندها، در زمانیکه مطابق رسم الخط باید جدا از ریشه کلمه ظاهر شوند، به همراه یک ZWNJ در ابتدای پسوند و یا در انتهای پیشوند رمزنگاری و نمایش داده می‌شوند. کاراکتر فاصله تنها در میان دو کلمه مستقل فارسی استفاده می شود. شرح کامل تر از این مطلب در [۳] و [۸] آمده است.



شکل ۱. اگر تمامی شکل را به عنوان استاندارد ارائه شده برای برچسب دهی پیکره در نظر بگیرید، آنگاه سیاست اتخاذ شده در توکن بندی، دسته بندی ارائه شده برای مقوله های گفتاری، و رسم‌الخط زبان، المانهای اساسی هستند که مستقیما بر روی مجموعه برچسب های تعریف شده برای حاشیه نویسی پیکره تاثیر می گذارند.

با مطالعات صورت گرفته بر روی زبان فارسی [۹][۱۰][۱۱][۱۲][۱۳] تقسیم بندی جدیدی از مقوله های فارسی جهت ارائه آن در چارچوب MULTEXT-East ارائه شده است. این دسته بندی جدید قائل به یازده مقوله گفتاری برای زبان فارسی است. جدول شماره یک این مقوله ها را به همراه تعداد ویژگی های در نظر گرفته شده برای آنها نمایش می دهد؛ برای هر یک از این مقوله‌ها ویژگی های صرفی-نحوی خاص در نظر گرفته شده است و در چارچوب MULTEXT-East ارائه شده است. به این دلیل پاره ای از ویژگی ها و یا مقادیر آنها صرفا جهت فارسی به این چارچوب افزوده شده است. همانطورکه برخی از ویژگی های از پیش تعریف شده برای دیگر زبان ها برای فارسی قابل کاربرد نیستند.

نظر نویسندگان مقاله و نتایج حاصل، بر این است که روش ارائه شده به همراه تعاریف جدید ارائه شده از مقوله های گفتاری فارسی و ویژگی های صرفی-نحوی آن بهترین نتیجه را در تحلیل رایانه ای فارسی به همراه دارد. در ادامه و در بخش بعد ویژگی های صرفی-نحوی ارائه شده برای فارسی و روش رمزنگاری آنها در چارچوب MULTEXT-East توضیح داده شده است.

جدول (۱). مقوله های گفتاری در فارسی به همراه تعداد ویژگی ها

Part of Speech	Code	Number of Attributes
Noun	N	4
Verb	V	10
Adjective	A	4
Pronoun	P	6
Determiner	D	1
Adverb	R	2
Adposition	S	2
Conjunction	C	2
Numeral	M	3
Interjection	I	0
Abbreviation	Y	0

۳- ویژگیهای صرفی-نحوی فارسی

روش ارائه ویژگی ها در چارچوب MULTEXT-East، روشی مبتنی بر مکان است و هر ویژگی تنها با یک کاراکتر نمایش داده می شود. در این چارچوب اولین کارکتر از یک حاشیه نویسی مقوله گفتاری واژه را مشخص می نماید. سپس مقادیر ویژگی ها با توجه به محل از پیش تعریف شده برای آن، در جلوی این کارکتر نوشته می شوند. در صورتیکه یک ویژگی برای یک زبان قابل تعریف نباشد، در مکان مرتبط با این ویژگی خط تیره (-) قرار می گیرد. جهت اطلاع بیشتر از روش ارائه ویژگی ها در این چارچوب می توان به [۱۴] مراجعه نمود.

در جداول پیش رو ویژگی های صرفی-نحوی فارسی برای هر مقوله آورده شده است. اولین ستون از سمت چپ در این جداول محل قرارگیری مقادیر آن ویژگی را در یک حاشیه نویسی نمایش می دهد. برای کوته نویسی، ویژگی های غیر قابل کاربرد برای فارسی از جداول حذف شده اند. بنابراین ممکن است در برخی جداول مقادیر برای برخی مکان ها تعریف نشده باشد، در این صورت در محل این ویژگی‌ها همواره یک خط تیره قرار می گیرد. ستون دوم در این جدول نام ویژگی را نمایش می دهد به عنوان مثال ویژگی شخص برای مقوله گفتاری فعل. ستون سوم و چهارم به ترتیب مجموعه مقادیر با معنی برای این ویژگی و کارکتر نمایش دهنده این مقدار را نمایش می دهند. به این ترتیب برای ویژگی شخص برای مقوله گفتاری فعل، سه مقدار با معنی شامل اول، دوم و سوم وجود دارد که به ترتیب با کارکترهای ۱،۲و ۳ در مکان چهارم پس از کارکتر نمایش دهنده مقوله گفتاری فعل (V) قرار می گیرند.

جداول ۲ تا ۱۰ ویژگی های صرفی-نحوی برای مقولات گفتاری مختلف، نحوه رمزنگاری، و مقادیر در نظر گرفته شده برای آنها را نمایش می دهد. این جداول برای دو مقوله گفتاری صوت و اختصارات، به دلیل در نظر گرفتن هیچ ویژگی صرفی-نحوی نیامده است.

جدول (۲). ویژگیهای صرفی-نحوی مقوله اسم

P	ATT	VAL	C
1	Type	common	c
		proper	p
3	Number	singular	s
		plural	p
4	Case	genitive	g
5	Definiteness	no	n
		yes	y

جدول (۳). ویژگیهای صرفی-نحوی مقوله فعل

P	ATT	VAL	C
1	Type	main	m
		auxiliary	a
		modal	o
		copula	c
		light	l
2	VForm	indicative	i
		subjunctive	s
		imperative	m
		participle	p
3	Tense	present	p
		past	s
4	Person	first	1
		second	2
		third	3
5	Number	singular	s
		plural	p
8	Negative	no	n
		yes	y
10	Clitic	no	n
		yes	y
14	Aspect	progressive	p
15	Courtesy	no	n
		yes	y
16	Transitive	no	n
		yes	y

جدول (۴). ویژگیهای صرفی-نحوی مقوله صفت

P	ATT	VAL	C
1	Type	qualificative	f
2	Degree	positive	p
		comparative	c
		superlative	s
5	Case	genitive	g
6	Definiteness	no	n
		yes	y

جدول (۵). ویژگیهای صرفی-نحوی مقوله ضمیر

P	ATT	VAL	C
1	Type	personal	p
		demonstrative	d
		indefinite	i
		interrogative	q
		reflexive	x
		reciprocal	y
2	Person	first	1
		second	2
		third	3
4	Number	singular	s
		plural	p
5	Case	genitive	g
		accusative	a
8	Clitic	no	n
		yes	y

جدول (۶). ویژگیهای صرفی-نحوی مقوله **Determiner**

P	ATT	VAL	C
1	Type	demonstrative	d
		indefinite	i
		interrogative	q
		exclamative	e
		article	a
		exceptional	x
4	Number	singular	s
		plural	p

جدول (۷). ویژگیهای صرفی-نحوی مقوله قید

P	ATT	VAL	C
2	Degree	positive	p
		comparative	c
7	Case	genitive	g

جدول (۸). ویژگیهای صرفی-نحوی مقوله Adpositions

P	ATT	VAL	C
1	Type	preposition	p
		postposition	t
2	Formation	simple	s
		compound	c

جدول (۹). ویژگیهای صرفی-نحوی مقوله حرف ربط

P	ATT	VAL	C
1	Type	coordinating	c
		subordinating	s
2	Formation	simple	s
		compound	c

جدول (۱۰). ویژگیهای صرفی-نحوی مقوله عدد

P	ATT	VAL	C
1	Type	cardinal	c
		ordinal	o
		fractal	f
		ordinal2	r
4	Case	genitive	g
6	Definiteness	no	n
		yes	y

شرح دقیق از ویژگی های ارائه شده و دلیل ارائه هر یک از این ویژگی ها در [۳] آمده است. جدول ۱۱ تعداد برچسب های مختلف با معنی برای هر مقوله گفتاری را نمایش می دهد. جمعاً امکان وقوع ۷۷۱ برچسب مختلف با معنی برای کلمات فارسی وجود دارد. از بین مقوله های گفتاری مختلف فعل ها بیشترین تعداد برچسب را به خود اختصاص داده اند.

جدول (۱۱). تعداد برچسب های با معنی برای هر مقوله گفتاری

Part of Speech	Number of Attributes
Noun	12
Verb	639
Adjective	12
Pronoun	78
Determiner	6
Adverb	3
Adposition	3
Conjunction	4
Numeral	12
Interjection	1
Abbreviation	1
Total Number	771

۴- ۴. پیکره

همانطورکه که گفته شد، در چارچوب MULTEXT-East رمان ۱۹۸۴ اثر جرج ارول به عنوان متن اصلی و پیکره متنی انتخاب شده است. بنابراین نسخه فارسی این کتاب نیز برای اضافه کردن زبان فارسی به این چارچوب حاشیه نویسی شده است. بطورکلی این پیکره از:

- ۱۱۰۰۰۰ توکن
- ۱۱۲۶۶ پاراگراف
- ۶۶۰۶ جمله
- ۶۶۳۲ لما
- ۱۳۵۹۷ نوع کلمه

تشکیل شده است. از میان ۷۷۱ برچسب مختلف با معنی و ممکن برای فارسی تنها ۴۴۸ برچسب مختلف در این پیکره رخ داده است.

این پیکره در قالب بسته ارائه شده در MULTEXT-East ارائه، و به صورت بر خط جهت تحقیق، به صورت رایگان در اختیار محققین و دانشجویان علاقمند به پردازش زبان طبیعی، قرار خواهد گرفت.

۵- نتیجه

این مقاله به معرفی اولین قدم ها جهت ساخت منابع اساسی جهت پردازش زبان فارسی می پردازد. در اولین قدم جهت حرکت به سوی این هدف، ویژگی های صرفی و نحوی بر ای زبان فارسی مطابق با خط مشی های ارائه شده توسط EAGLE/MULTEXT ارائه شده است. کار انجام شده، دسته بندی جدیدی از مقوله های گفتاری زبان فارسی به همراه ویژگی های صرفی-نحوی جدید، معرفی می نماید. علاوه بر این، مقاله ارائه شده، روش جدید جهت ارائه متون الکترونیکی فارسی ارائه می نماید. به طور کلی روش انتخاب شده جهت ارائه متون و همچنین دسته بندی ارائه شده از مقوله های گفتاری در جهتی صورت گرفته است که نخست همگی ویژگیهای صرفی-نحوی زبان فارسی را

کاملا و به شکل واقعی ارائه نماید و در عین حال حداکثر هم خوانی و سازگاری با دیگر زبانها را داشته باشد.

در قدم بعدی جهت آماده سازی منابع اساسی برای فارسی، پیکره ۱۹۸۴ به فارسی، بر چسب دهی شده است. وجود این پیکره برای زبان فارسی علاوه بر فراهم آوردن یک مجموعه داده استاندارد جهت استفاده محققین پردازش زبان طبیعی در فارسی، امکان استفاده از متدهای آماری جهت پردازش فارسی را معرفی می نماید. علاوه بر این چند زبانه بودن این پیکره، سبب میشود تا بتوان از آن در تحقیقات مرتبط با ترجمه ماشینی از فارسی/ به فارسی از دیگر زبانهای موجود در این چارچوب، و در نهایت تحقیق و گسترش چنین سامانه هایی پرداخت. علاوه بر اینکه وجود زبان فارسی در یک چارچوب استاندارد و موازی با دیگر زبانها، امکان مطالعه و تحقیق جهت استخراج دانش به کمک پردازش های بین زبانی و یا حتی مقایسه بین این زبان ها و زبان فارسی، فراهم می آورد.

به عنوان کار آینده، تصمیم بر گسترش پیکره به ۱ ملیون کلمه می باشد. علاوه بر اینکه گسترش سیستمی جهت استاندارد سازی خودکار متون فارسی نیز مد نظر قرار گرفته است.

سپاسگزاری

نویسندگان این مقاله مایلند از پرفسور توماژ اریاوتز[4] و پرفسور دمیر چوار[5] به دلیل هم‌فکری و راهنمایی های بسیارشان تشکر نمایند.

مراجع

[1] Strik, H. Daelemans, W. Binnenpoorte, D. Sturm, J. de Vriend, F. and Cucchiarini, C.: *Dutch HLT resources:From BLARK to priority lists*, In Proceedings of ICSLP, Denver, USA, pp. 1549-1552, Denver, USA, (2002).

[2] Krauwer, S. Maegaard, B. Choukri, K. and Damsgaard Jørgensenm, L.: *Report on BLARK for Arabic*, (2004).

[3] QasemiZadeh B., Rahimi S.: *Persian in MULTEXT-East Framework*. FinTAL 2006: 541-551, Springer Publisher, Lecture Notes in Computer Science, Vol. 4139, 2006.

[4] Ide N. and Veronis J.: Multext: Multilingual Text Tools *And Corpora*. In 15th Int. Conference On Computational Linguistics, Pages 588–592, Kyoto, Japan, (1994).

[5] Erjavec T., Krstev C., Petkevic V., Simov K., Tadic M., and Vitas D.: *The MULTEXT-East Morphosyntactic Specifications For Slavic Languages*, Proceedings Of The EACL 2003 Workshop On The Morphological Processing Of Slavic Languages, (2003).

[6] Isiri 6219:2002: Information Technology – Persian Information Interchange and Display Mechanism, Using *Unicode*, (2002).

[7] Iran's Academy Of Persian Language and Literature: *Official Persian Orthography*, ISBN: 964-7531-13-3, 3rd Edition, (2005).

[8] QasemiZadeh B., Rahimi S., Safaee M., *Challenges in Persian Electronic Texts Analysis*, InScit2006, Merida, Spain, 2006.

[9] Hasan A. and Ahmadi Givi H.: Persian Grammar, ISBN964-318-007-7, 22nd Edition, Tehran, (2002).

[10] Meshkatodini M.: Introduction to Persian *Transformational Syntax*, 2nd Edition, ISBN: 964-6335-80-2, Ferdowsi University Press, (2003).

[11] Lazard, G.: A Grammar of Contemporary Persian, Mazda Publishers, (1992).

[12] Riazati D.: Computational Analysis of Persian Morphology, Msc Thesis, Department Of Computer Science, RMIT, (1997).

[13] Bateni, M.: Towsif-E Sakhteman-E Dastury-E Zaban-E Farsi [Description Of The Linguistic Structure Of Persian Language], Amir Kabir Publishers, Tehran, Iran, (1995).

[14] Erjavec T.: MULTEXT-East Morphosyntactic *Specifications*, Version 3.0. Supported By EU Projects Multext-East, Concede And TELRI, (2004).

زیرنویس‌ها

[1] Basic language resources kits (BLARK)

[2] The Hub language

[3] Morphosyntactic Descriptions

[4] Tomaz Erjavec

[5] Damir Cavar