بررسی نظام واژه‌سازی به روشِ تجزیهٔ تکواژی صوری

کاترین کریکونیوک (دانشجوی دکترای زبان‌شناسی در انستیتوی کریمسکیی، کی‌یف، اوکراین)

سیستم امروزی واژه‌سازی زبان فارسی نتیجهٔ تحوّل پیچیدهٔ تاریخی واژگان فارسی است. رایج‌ترین نوع فرایند واژه‌سازی در زبان فارسی (افزودن پیشوند و پسوند) مختصّ همهٔ زبان‌های هند و اروپایی است. امّا، در زبان فارسی، فرایند افزودن پسوند غالب است. زبان فارسی، به‌جز واژگانی که از زبان‌های ایرانی به ارث برده است، دارای لایه‌هایی گرانبار از واژه‌های به وام گرفته از زبان‌های عربی و ترکی و فرانسه و یونانی و جز آن است که، به شیوه‌های گوناگون، نظام واژه‌سازی زبان فارسی را تحت تأثیر قرار داده‌اند. با این حال، اغلب اوقات، ارزیابی همبستگی وام‌واژه‌ها و واژه‌های اصیل به صورت حسّی انجام می‌گیرد و نیاز به پژوهش دقیق‌تری دارد.

۱. معرّفی روش تجزیهٔ تکواژی صوری

تحقیق در واژه‌سازی، گامی الزامی در راستای ایجاد تصوّر کلّی از ساختار هر زبانی است. پیچیدگی نظام واژه‌سازیِ هر زبانی از رابطهٔ تنگاتنگ آن با سطوح ساختاری دیگر زبان و انعطاف‌پذیریِ نظام و عدم امکان تمایز ساختِ واژه (word formation) و صرفِ واژه (inflection) همچنین او تنوّع و اهمّیّت ارتباطی کارکردهای نظام ساخت واژه ناشی می‌شود.

دانشمندان بسیاری، مستقیم یا غیر مستقیم، موضوع ساخت واژه در زبان فارسی را مطالعه کرده‌اند. از پژوهشگران روسی و غربی، روبینچیک[۱]، پیسیکوف[۲]، اُوچینُیکُوا[۳]، بِرتِلس[۴] و میس لَمْتُن[۵] را می‌توان نام برد. زبان‌شناسان ایرانی از قبیل پرویز خانلری، خسرو فرشیدورد، علی‌اشرف صادقی، علاءالدّین طباطبائی، و رضائی باغ‌بیدی به تحقیق درواژه‌سازی زبان فارسی پرداخته‌اند. برخی از نکات تکواژشناسیِ فارسی در پژوهش‌های زبان‌شناسیِ رایانه‌ای کارِن مِگِردومیان، محسن عرب‌سرخی، مهرنوش شمس‌فرد، و ساگوت و والتِر[۶] منعکس شده‌اند. امّا، با وجود توصیف‌های گرامِری متعدّدِ زبان فارسی، تا به حال نظام ساخت واژگان فارسی به حیث کُلِّ چند بعدی و منشعب بررسی نشده است.

یکی از اوّلین تلاش‌ها در زمینهٔ تحقیق دربارهٔ نظام واژه‌سازی زبان فارسی با استفاده از روش تکواژی صوری، به همّت بیدوف[۷]، زیر نظر پروفسور تیشچنکو[۸]، در پایان‌نامه‌ای با عنوان «ساختار تکواژی زبان فارسی»[۹] بر اساس واژه‌نامهٔ فارسی ـ ژاپنیِ اُکازاکی (حدود شش هزار واژه)، صورت گرفت.

تکواژشناسیِ صوری، در سال‌های شصت قرن پیش، در چارچوب زبان‌شناسی ساختگرای روسی، پدید آمد. یکی از روش‌های مهمّ آن تجزیهٔ تکواژیِ صوری است که اصول آن را براتچیکوف[۱۰]، فیتیالوف[۱۱]، و تسیتین[۱۲]، به منظور ترجمهٔ ماشینی در زمینهٔ صرف واژه، معرّفی کرده‌اند. در سال‌های ۱۹۶۳-۲۰۰۳، تیشچنکو، با اختیار روش تجزیهٔ تکواژی صوری، تحقیق در نظام صرف واژهٔ زبان‌های گوناگون را، با مروری بر نظام‌های زیرمجموعه‌ای صرف فعل در زبان‌های فرانسه و ایتالیایی و پرتغالی و اسپانیائی و فارسی و فنلاندی، انجام داد و آن پژوهشی بود در الگوهای تکیه[۱۳] درواژه‌سازیِ اسامی زبان روسی همچنین تحلیل الگوهای صرفِ واژه‌های پُرکاربُرد اوکراینی.

---

1) Yu. A. Rubinchik 2) L. S. Peisikov 3) I. K. Ovchinnikova
4) E. Ye. Bertels 5) A. K. S. Lambton 6) B. Sagot & G. Walter
7) O. Bedov 8) K. M. Tyschenko
9) "Morphological structure of the Persian language"
10) I. Bratchikov 11) S. Firialov 12) G. Tseitin 13) accentuation models

روش تجزیهٔ تکواژی صوری مؤثّرترین روش شناخت ساختار واژه شمرده می‌شود (→ 5,16) که امکان به دست آوردن اطّلاعاتی دربارهٔ واژه از راه تقطیع دقیق آن و دسته‌بندی همزمان اطّلاعات به دست آمده را فراهم می‌سازد (→ Ibid, 25). امکانات ترکیب‌بندی تکواژها (در زبان‌شناسی همزمانی) همیشه محدود است. از این رو، انتظار می‌رود که روند تجزیهٔ تکواژی مراحلِ معدودی را طی کند (→ Ibid, 16). به علاوه، جنبهٔ صوریِ این روش اجازه می‌دهد که نسبت‌ها میان سازه‌های واژه اجمالاً نشان داده شود و ماهیّت آن نسبت‌ها ارزیابی و ساختار منطقیِ کلّ سیستم مشخص گردد.

در تجزیهٔ تکواژی صوریِ[۱۴] هر سیستم زبانی، مجموعه‌ای از اصول روش‌شناختیِ اولیّه به‌کار می‌رود که شامل **تمامیّت**، **سادگی**، **همریختی**، و **صوری بودن** تحقیق است.

تمامیّتِ تجزیهٔ تکواژی، با استقرای همهٔ پدیده‌های زبان در حوزهٔ موضوع بررسی در مرحلهٔ اولیّهٔ تحقیق، تأمین می‌شود که در یک یا چند منبع درج شده است.

سادگی در تجزیهٔ تکواژی با انتخاب ابزار سادهٔ مطالعهٔ زبان و شفّاف‌سازیِ روند تجزیه و تحلیل حاصل می‌شود و بررسی مجدّد کامل و بدون مغایرت با پژوهشِ صورت‌گرفته را امکان‌پذیر می‌سازد. در این مورد، باید مراقب بود که ابزار مطالعهٔ زبان از خودِ موضوع تحقیق پیچیده‌تر نباشد.

در فرمول‌بندی ریاضی‌وار، استفاده از دستگاه نشانه‌های صوری برای تکواژها و مقوله‌های دیگر واژه‌سازی ضرورت دارد. شیوهٔ سنّتی برای نشان دادن مشخّصات گرامریِ هر واژه پذیرفته نیست.

همریختی با اجتناب از تلفیق التقاطیِ ابزار متناقضی حاصل می‌شود که در چارچوب روش‌های دیگری پدید آمده‌اند. همریختی همچنین به انتخاب موضوع تحقیق مربوط می‌شود که مستلزم توجّه تمام به گروه پدیده‌هایی است که وقوع[۱۵] آنها بیشتر است و نه پدیده‌های دارای وقوع نادر. (→ 3, 11)

استفاده از نشانه‌هایی به منظور ترسیم نظام کلّی واژه‌سازی در زبان فارسی تجزیهٔ ترکیب‌ها را امکان‌پذیر می‌سازد. فرمول نشانه‌گذاری واژه‌ها به صورت زیر است:

14) formalized morphological description 15) occurence

$$R''=mR'\ m_1mR'_2...,$$

در این فرمول، R رادیکال (ریشه) یا واژه‌ای تجزیه‌ناپذیر در یکی از مقولات صرفی است که با نشانه‌های Aj (صفت) و Verb (فعل) و Ad (قید) و N (اسم) و جز آن علامت‌گذاری می‌شود؛ m پیشوند یا پسوند است که عیناً با حروف لاتینی نوشته می‌شود. علامت ثانیه second (″) به این معناست که واژه ترکیبی است و از یک یا چند رادیکال (ریشه) و یک یا چند وند تشکیل می‌شود. (2→)

کلمات مرکّب، به ترتیب، به کوچک‌ترین واحدهای معنی‌دار و دارای نقش دستوری (تکواژ) تجزیه می‌شوند. با این فرمول‌بندی، رهایی از معنی واژه و تمرکز روی ساختار میسّر می‌گردد. به عنوان مثال:

بهبودی behbudi-N'' = Aj+ W*+ i

اعجمی mi a*jami-Aj'' = Aj+ i

پسرانه pesarâne-Ad'' = N+ âne

خندیدن xandidan-$Verb''$ = W+ an

از این مثال‌ها پیداست که، در قدم بعدی، الگوهای واژه‌سازیِ[16] به دست آمده بدون هیچ مشکلی گروه‌بندی می‌شوند که برای فهمیدن ساختار کلّ نظام واژه‌سازی در زبان فارسی حایز اهمّیّت است.

۲. قواعد زبان‌شناختی در نظام واژه‌سازی زبان فارسی امروز

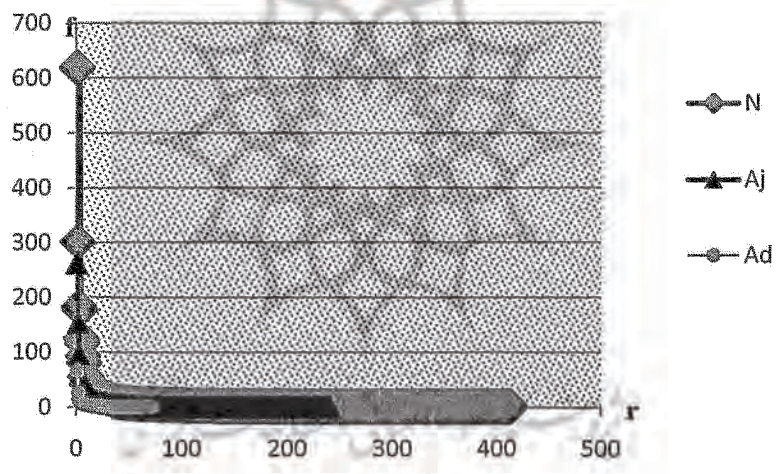
با کاربُرد روش‌های آماری در زبان‌شناسی، کسب نتایج دقیق از پدیده‌های موضوع بررسی و مطالعهٔ روابط آنها میسّر می‌گردد.

قانون استو- زیپْف[17] ناظر به نسبت تجربی توزیع کلمات در همهٔ زبان‌هاست با توجّه به بسامد آنها به ترتیب از بالا به پایین. امروزه این نسبت را با نام **قانون زیپف** می‌شناسند، هرچند زیپف تنها اظهارنظرهای اِستو و کُندُن[18] را کامل کرد (75 , 4→). زیپف، در اثر خود،

16) word formation patterns
17) Estoup-Zipf's law
18) Estoup J. B., E. Condon

زیست‌شناسی روان[19] (۱۹۳۵)، نشان داده است که بسامد واژه با رتبهٔ آن (شمار تکواژهای سازندهٔ آن) نسبت معکوس دارد (89 , 1←). وی، در فرمول استو-کُندَن، ثابت Y را وارد کرد (rf^{Y}=const) که باید برای هر متنی محاسبه شود. زبان‌شناسان و ریاضی‌دانان دیگری از قبیل گیرو، ایُول، مَندِلبرُت، هِردَن، برادْفورد، هِرتْز[20] این فرمول را کامل‌تر کردند. فرمول نهائی با افزودن چند متغیّر به دست آمده و با داده‌های تجربی مطابقت بیشتری کرده است. قانون **استو- زیپف**، مانند همهٔ مدل‌های آماری، ناظر به نزدیک شدن به واقعیّت است. در نمودار، این نسبت به صورت منحنی هذلولی درمی‌آید. (22 , 6←)

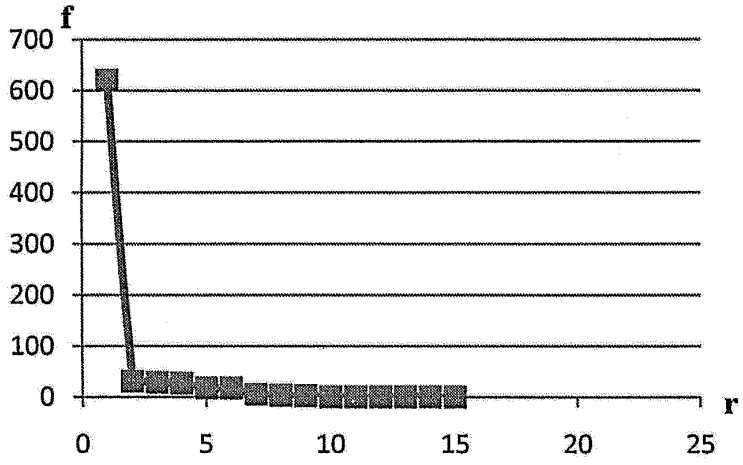
تجزیهٔ تکواژیِ صوری بر روی واژه‌های واژه‌نامهٔ نوین محمّد قریب (۲۹۱۶۲ واژه) کارآمدیِ قانون **استو- زیپف** را در ساختار نظام واژه‌سازی زبان فارسی نمودار ساخت یعنی نشان داد که، هرچه رتبهٔ عناصر (تکواژها) در واژه‌سازی بالاتر باشد، بسامد آنها کمتر است.



شکل ۱- قانون استو- زیپف ۱: توزیع رتبه‌ای- بسامدیِ مدل‌های واژه‌سازی اسم‌ها و صفت‌ها و قیدهای ترکیبی. f نشانهٔ شمارواژه‌هایی است که در مدل واژه‌سازی مندرج است. r رتبهٔ مدل واژه‌سازی است.

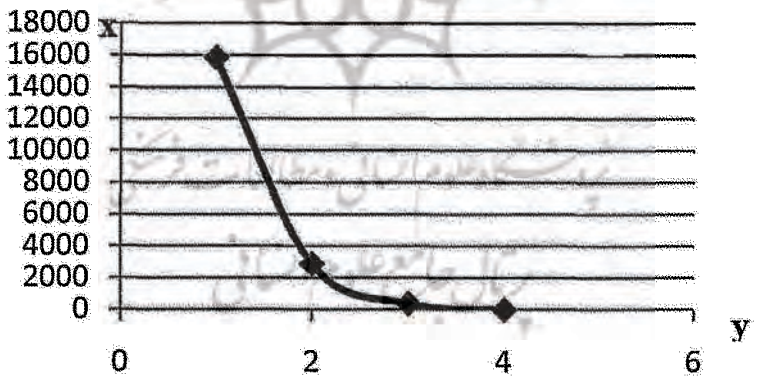
19) *The Psychobiology of Language*. An Introduction to Dynamic Philology. Boston, 1935.

20) P. Guiraud, G. Udny Yule, B. Mandelbrot, G. Herdan, S. C. Bradford, M. B. Arapov, Hertz.

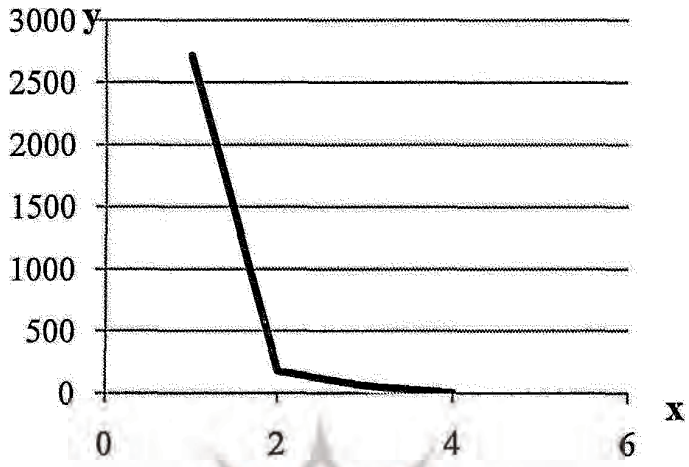


شکل ۲ـ قانون استو ـ زیپف ۱: توزیع رتبه‌ای ـ بسامدی مدل‌های واژه‌سازی فعل‌های ترکیبی.
f نشانهٔ شمار واژه‌هایی است که در مدل واژه‌سازی مندرج است؛ r نشانهٔ رتبهٔ مدل واژه‌سازی است.

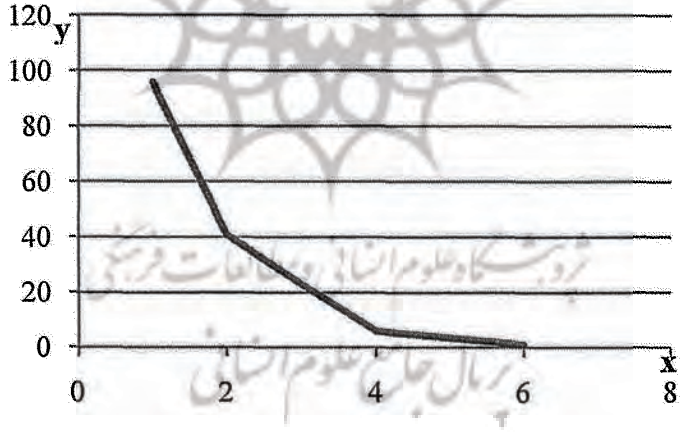
توزیع مدل‌های واژه‌سازی با در نظر گرفتن شمار تکواژها در آنها کارآمدی قانون **استو ـ زیپف** 2 را نمودار ساخت یعنی نشان داد که هرچه شمار تکواژها در مدل‌های واژه‌سازی کمتر باشد، بسامد شمار مدل‌ها (ترکیبات) بیشتر می‌شود.



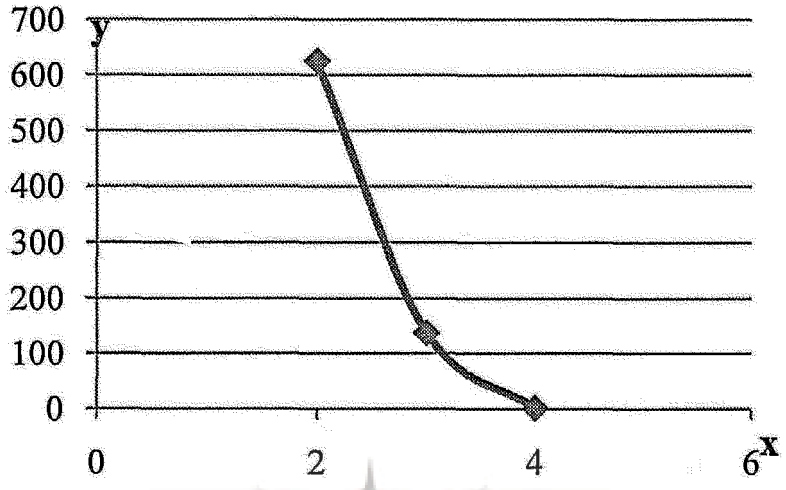
شکل ۳ـ قانون استو ـ زیپف ۲: توزیع مدل‌های واژه‌سازی اسم با در نظر گرفتن شمار تکواژها.
x نشانهٔ شمار واژه‌ها؛ y نشانهٔ شمار تکواژها در مدل‌های واژه‌سازی است.



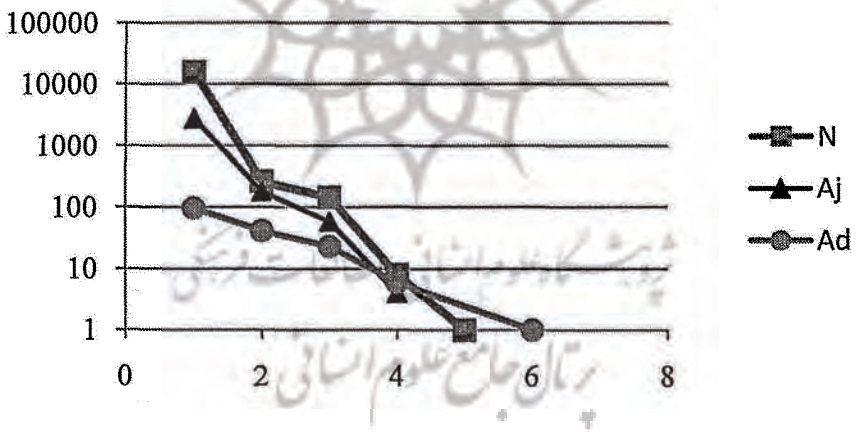
شکل ۴ـ قانون استو ـ زیپف 2 توزیع مدل‌های واژه‌سازی صفت با در نظر گرفتن شمار تکواژها.
x نشانهٔ شمار واژه‌ها؛ y نشانهٔ شمار تکواژها در مدل‌های واژه‌سازی است.



شکل ۵ـ قانون استو ـ زیپف 2 توزیع مدل‌های واژه‌سازی صفت با در نظر گرفتن شمار تکواژها.
x نشانهٔ شمار واژه‌ها؛ y نشانهٔ شمار تکواژها در مدل‌های واژه‌سازی است.



شکل ۶ـ قانون استو ـ زیپف 2: توزیع مدل‌های واژه‌سازی فعل با در نظر گرفتن شمار تکواژها.
x نشانهٔ شمار واژه‌ها؛ y نشانهٔ شمار تکواژها در مدل‌های واژه‌سازی است.



شکل ۷ـ قانون استو ـ زیپف 2: توزیع مدل‌های واژه‌سازی اسم و صفت و قید با در نظر گرفتن شمار تکواژها. (مقیاس لگاریتمی)
x نشانهٔ شمار واژه‌ها؛ y نشانهٔ شمار تکواژها در مدل‌های واژه‌سازی است.

نتیجه

در این مقاله، روش نوین بررسی نظام واژه‌سازی در زبان فارسی معرّفی شده است. تجزیهٔ تکواژیِ صوری یکی از مؤثّرترین روش‌های شناخت ساختار واژه است که کسب اطّلاعاتی دربارهٔ واژه را از راه تجزیهٔ آن و دسته‌بندی اطّلاعات به دست آمده میسّر می‌سازد. طیّ تجزیهٔ تکواژیِ صوریِ هر دستگاه زبانی، مجموعه‌ای از اصول روش‌شناختی اوّلیّه به‌کار می‌رود که **تمامیّت**، **سادگی**، **همپیوستگی**، و **خصلت صوریِ تحقیق** را تضمین می‌کنند. کاربُرد تجزیهٔ تکواژیِ صوری کارآمدی قانون استو- زیپف را در مطالعهٔ ساختار نظام واژه‌سازی زبان فارسی آشکار ساخته است.

منابع

1. Creutz, M. (2006). *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmention with Aplication to Automatic Speech Recogrition*. Dissertation for the Degree of Doctor of Science in Technology, Helsinki University of Technology (Espoo, Finland), p. 110.

2. Бєдов О. Морфологічна структура сучасної перської лексики (дипломна робота) / Олександр Бєдов. - Київ: 1997. (Препринт).

3. Исследования в области вычислительноиˇ лингвистики и лингвостатистки. Сборник. Ответственныйˇ редактор_ доцент В. М. Андрющенко. М., Изд-во МГУ, 1978. _ 191 с.

4. Орловˇ. К. Невидимая гармония. Сб. «Число и мысль». Вып. 3. _ М.: «Знание», 1980. _ С. 70-105.

5. Тищенко К. М. Глагольная парадигма романских языков/ дис. на соиск. уч. ст. канд. филолог. наук / К. Н. Тищенко. - Киев, 1969. - 160 с.

6. Тищенко К. М. Основи мовознавства: Системнийˇ підручник / Костянтин Миколайович Тищенко. _ Київ_ 308 с.: Видавничо-поліграфічний центр «Київський університет», 2007.

کتابنامه

Dolamic, L. & Savoy, J. (2009). *Persian Language, is Stemming Efficient?* Available at:http://www.uni-weimar.de/medien/webis/research/events/tir-09/tir09-papers-final/dolamic09-persian-language-is-stemming-efficient.pdf.

Lambton, A. K. S. (2003). *Persian Grammar*. Cambridge: Cambridge University Press, p. 332.

Meqerdoomian, K. (2000). *Unification-based Persian morphology.* Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.27.6355&rep=rep1&type=pdf.

Овчинникова, И., Мамед-заде, А.К. Учебник персидского языка. (Часть 1). - Москва: Издательство Московского Университета, 1966. - 444 с.

Бертельс Е. Э. Грамматика персидского языка. - Ленинград: Издание Института живых восточных языков им. А. С. Енукидзе, 1926. - 127 с.

Братчиков И. Л., Фитиалов С. Я., Цеитин Г. С. О структуре словаря и кодировке информации для машинного перевода. // В кн.:Материалы по машинному переводу. – Л.: Изд-во ЛГУ, 1958. - с. 61-87.

Конверський А. Є. Логіка (традиційна та сучасна): Підручник для студентів вищих навчальних заклад ів. - Київ: Центр учбової літератури, 2008. – 536 с.

Лингвистический энциклопедический словарь // Под ред. В. Н. Ярцевой. –Москва: «Советская энциклопедия», 1990. - 685 с.

Рубинчик˙. А. Грамматика современного персидского литературного языка. – М.: Издательская группа "Восточная литература" РАН, 2001. – 600 с.

Рубинчик˙. А. Предисловие // Персидско-русский словарь / [ред. Рубинчик˙. А.]. В 2-хт. – Т.1. – М.: Сов. Энциклопедия, 1970. – С. 5-16.

Тищенко К. М. Оптимальна морфологія перського дієслова // Вісник КДУ. Іноз. Філологія. Вип. 25. – К.: КДУ, (о, з, %, є.).

Тищенко К. М. Службові дієслова у підсистемі дієслів перської мови // К. Н. Тищенко, А. Півторак // III сходознавчі читання А. Кримського. Тези міжнар. наук. конф. - К., 1999.

Фитиалов С. Я. О построении формальной морфолгии в связи с машинным переводом: доклады на конференции по обработке информации, машинному переводу и автоматическому чтению текста. – Москва: Производственно-издательский комбинат ВИНИТИ, 1961. – 24 с.

Фрумкна Р. М. Роль статистических методов в современных лингвистических исследованиях // Математическая лингвистика. – М.: Изд-во «Наука», 1973. – С. 165-182.

Чурсин, Н. Н. Популярная информатика. – К.: Техника, 1982. – 158 с.

فرشیدورد، خسرو (۱)، فرهنگ پیشوندها و پسوندهای زبان فارسی (همراه گفتارهایی دربارهٔ دستور زبان فارسی)، انتشارات زوّار، تهران ۱۳۸۶.
ـــــ(۲)، دستور مفصّل امروز بر پایهٔ زبان‌شناسی جدید: شامل پژوهش‌های تازه‌ای دربارهٔ آوا‌شناسی و صرف و نحو فارسی معاصر و مقایسهٔ آن با قواعد دستوری، انتشارات سخن، تهران ۱۳۸۲.
قریب، محمّد، واژه‌نامهٔ نوین، تهران ۱۳۴۶.
لغت‌نامه دهخدا، ۱۵ جلد، مؤسّسهٔ انتشارات و چاپ دانشگاه تهران ۱۳۷۷.

❑