[25] Delimiter
[26] Multi-Unit Token(MUT)
[27] Compositional Construction
[28] Non-Compositional
[29] Semi-Compositional
[30] Compound
[31] Template
[32] Complex Predicate(CPr)
[33] Non-Verbal(NV)
[34] Light Verb(LV)
[35] idiom
[36] Distributed Morphology
[37] inchoative
[38] vP-shell
[39] Predicative Noun
[40] Predicative Adjective
[41] MTU Verbal Template
[42] MTU Prepositional Template
[43] MTU Conjunctional Template
[44] MTU Adverbial Template
[45] MTU Adjectival Template
[46] MTU Nominal Template

Department of Computer and Information Science, University of Pennsylvania (1991). ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz
[4] A. Bies and M. Maamouri "Penn Arabic Treebank Guidelines", Linguistic Data Consortium. University of Pennsylvania, 2003.
[5] K. Sima'an, A. Itai, Y.Winter, A. Altman and N. Nativ "Building a Treebank of Modern Hebrew Text", Traitment Automatique des Langues, 42:347–380, 2001.
[6] S.M Assi and M.H. Abdolhosseini "Grammatical Tagging of a Persian Corpus", The International Journal of Corpus Linguistics, Vol. 5, No.1, pp. 69-81, 2000.
[7] G. Leech "Corpus Annotation Schemes" Literary and Linguistic Computing", Vol. 8, No. 4, pp. 275-281, 1993.
[8] Jan Cloeren "In Syntactic Wordclass Tagging", In Hans van Halteren. Dordrecht: Kluwer Academic Publishers, 1999.
[9] G. Leech, and A. Wilson "EAGLES Recommendations for the Morphosyntactic Annotation of Corpora", EAGLES-Guidelines EAG--TCWG--MAC/R. Final version of 3.1996. EAGLES, Istituto di Linguistica Computazionale, Pisa, 1996.
[10] M. Bijankhan, J. Sheykhzadegan and M. Bahrani "Lessons From Designing A Persian Resource: Peykare" (under review)
[11] G. Lazard "La Formation de la Langue persane", Diffusion Peeters, Paris, 1995.
[12] K. Megerdoomian "Unification-Based Persian Morphology", In Proceedings of CICLing, 2000.
[13] W. L Graff "The Word and the Sentence", in Language, vol. 5, no. 3, pp. 163-188, Linguistic Society of America, 1929.
[14] P. L Garvin, , "Delimitation of Syntactic Units", in Language, vol. 30, no. 3, pp. 345-348, Linguistic Society of America, 1954.
[15] N. Chomsky and M. Halle "The sound pattern of English". New York: Harper and Row, Boston: MIT Press, 1968.
[16] P. Cole, J. Segui and M. Taft. "Words and morphemes as units for lexical access", Journal of memory and Language, 37, 312-330, 1997.
[17] Persian Academy of Language and Literature, Persian Orthography. 2005. http://www.persianacademy.ir/fa/dastoorpdf.aspx
[18] T. Buckwalter "Issues in Arabic Orthography and Morphology Analysis", In Proceedings of COLING, 2004
[19] J. Mohammad and S. Karimi "Light verbs are taking over: Complex verbs in Persian", *Proceedings of WECOL, pp. 195-212,* 1992.
[20] S. Karimi "A minimalist approach to Scrambling", Mouton de Gruyter, Berlin/New York, 2005.
[21] R. Folli, H. Harley and S. Karimi, "Determinants of event type in Persian complex predicates", *Lingua,* 2005.
[22] K. Hale and S. J. Keyser, "On argument structure and the lexical expression of syntactic relations", In: Hale, K., Keyser S.J., (Eds.), View from Building 20. MIT Press, Cambridge, MA, pp. 53-109, 1993.
[23] Gh. Karimi-Doostan, "Lexical categories in Persian", Manuscript submitted to the Journal of Linguistics for publication
[24] S. Karimi, " Persian Complex Verbs: Idiomatic or Compositional", Lexicology 3, 273-318, 1997
[25] A. Bies, M. Ferguson, K. Katz and R. MacIntyre, "Bracketing Guidelines for Treebank II Style: Penn Treebank Project", 1995. ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual
[26] V. Samiian, "Prepositions in Persian and the Neutralization Hypothesis", California State University, Fresno, 1991.
[27] Z. Abolhassani, "An Account for Compound Prepositions", Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 113–119, Sydney, July 2006.

مسعود شریفی آتشگاه دارای مدرک کارشناسی مهندسی کامپیوتر از دانشگاه صنعتی اصفهان ۱۳۷۶ و کارشناسی ارشد زبان‌شناسی از دانشگاه تهران ۱۳۷۹، درحال‌حاضر دانشجوی دکتری زبان‌شناسی دانشگاه تهران در مرحلهٔ دفاع از پایان‌نامه در زمینهٔ بانک درخت فارسی و علاقه‌مند به زبان‌شناسی پیکره‌ای و رایانه‌ای در زمینهٔ پردازش گفتار و متن می‌باشد.

محمود بی جن خان مدرک کارشناسی خود را در رشته ریاضی در سال ۱۳۶۰ از دانشگاه ایالتی تگزاس در آرلینگتن آمریکا دریافت کرد. او موفق به اخذ درجات کارشناسی ارشد و دکتری در رشته زبانشناسی از دانشگاه تهران به ترتیب در سال های ۱۳۶۹ و ۱۳۷۴ گردید. بی جن خان از سال ۱۳۷۲ به هیات علمی دانشکده ادبیات و علوم انسانی دانشگاه تهران پیوست. او در یازدهمین جشنواره بین المللی خوارزمی بخاطر ارائه طرح دادگان گفتاری زبان فارسی (فارس دات) حائز رتبه سوم تحقیقات کاربردی شد. زمینه های تحقیقاتی مورد علاقه وی آواشناسی واجشناسی زبانشناسی پیکره ای و رایانه ای است.

واژه‌نامه:

[1] **Treebank**
[2] **Text Segmentation**
[3] **Part of Speech (POS) Tag**
[4] **Dynamic**
[5] **Recoverability**
[6] **Consistency**
[7] **Peykare**
[8] **Multi-Token Unit (MTU)**
[9] **Static**
[10] **EAGLES Guidelines**
[11] **Grammatical mark-up**
[12] **Mnemonic**
[13] **Adposition**
[14] **Perso-Arabic**
[15] **Impure Abjad**
[16] **Diacritic**
[17] **Isolated**
[18] **Initial**
[19] **Medial**
[20] **Final**
[21] **Space**
[22] **Zero Width Non-Joiner(ZWNJ)**
[23] **Orthographic Variation**
[24] **Orthographic Token**

یکی دیگر از الگوهای صفتی واحدهای چندقطعه‌ای که بسیار نیز زایاست و در بخش ۳-۲ نیز به آن اشاره شد، همنشینی قطعه‌ها در آرایش نیمه‌ترکیبی است که در آن پیشوندهای عددی به همراه مقولهٔ اصلی با برچسب **متفرقه،صفت‌ساز**، تولیدکنندهٔ صفت مرکب هستند. مانند *سه نفره، دو زبانه، ۲۰ کیلویی، پنج لیتری*. ترکیبات **«عدد اصلی»** و **«متفرقه،صفت‌ساز»** همچنین می‌توانند به عنوان گروه قیدی با نقش معنایی حالت نیز ظاهر شوند که در بخش قبل به آن اشاره شد.

دسته‌ای از صفت‌ها در زبان فارسی از افزودن پسوند صفت‌ساز «-i» به مصدر ایجاد می‌شوند مانند دیدنی، شنیدنی. در مورد مصدرهای مرکب که جزء مصدری دارند نیز این مساله وجود دارد ولی چون در پیکره قطعه‌ها به صورت مجزا برچسب گرفته‌اند، لذا اینگونه صفت‌های چندقطعه‌ای نیز باید شناخته شده و برچسب واحد صفت بگیرند.مانند *از بین رفتنی*.

این مسأله در مورد افعال مرکب نیز با فراوانی بیشتر وجود دارد. به این صورت که فعل سبک به صورت صفت مفعولی نوشته می‌شود مانند *از حال رفته، پیش پا افتاده، از جان گذشته*.

دستهٔ وسیع دیگری از صفت‌های مرکب، تشکیل‌شده از قطعه‌های**«قابل»** و یا **«غیر قابل»** در نقش پیشوند به‌علاوهٔ اسم هستند مانند: *غیر قابل باور، قابل اعتماد*. در این گونه ترکیبات، قطعه‌های «قابل» و «غیر قابل» که به‌اشتباه، برچسب‌های جداگانهٔ صرفی گرفته‌اند، همانند پیشوندهای منفی‌ساز « **un-, in-**» و پسوندهای صفت‌ساز «**-able, -ible** » در انگلیسی هستند. ازسویی دیگر اگر این قطعه‌ها سرهم نوشته شده باشند در پیکره برچسب «صفت» گرفته‌اند.

بنابراین دلایل، پیشنهاد می‌شود که این قطعه‌ها نیز به عنوان صفت‌های مرکب در نظر گرفته شده و برچسب «صفت» بگیرند.

«در حال» به‌علاوهٔ **«اسم»**، نیز دستهٔ دیگری از صفت‌های مرکب را می‌سازند مانند: *در حال حرکت، در حال احداث*.

چنین قطعه‌های نیز بهتر است به عنوان یک صفت مرکب در نظر گرفته شده و برچسب «صفت» بگیرند.

۵-۵- اسم‌ها

همانگونه که در بخش ۴-۵ دیدیم اکثر واحدهای چندقطعه‌ای که سازندهٔ اسم هستند در پیکره به درستی برچسب واحدی گرفته‌اند.

در این بخش به «الگوهای اسمی واحدهای چندقطعه‌ای[۴۶]» اشاره می‌کنیم. همانگونه که در بخش ۴-۵ نیز دیدیم، دسته‌ای از اسم‌های مرکب از ترکیب دو اسم حاصل می‌شوند که برخی از این ترکیبات می‌تواند بسیار پویا باشد مثلاً اگر قطعهٔ دوم، «خانه» یا «فروش» باشد.

ترکیب پیشوند «شبه» با اسم‌های دیگر، سازندهٔ اسم مرکب است مانند: *شبه دارو، شبه وبا، شبه جزیره* . این نوع قطعه‌ها نیز که زایایی دارند در پیکره به عنوان یک اسم در نظر گرفته شده‌اند.

دستهٔ دیگر اسم‌های مرکب، تشکیل شده از یک حرف‌اضافه مانند «زیر،بالا، پیش، پس»به عنوان پیشوند و یک اسم و یا ستاک گذشته یا حال است مانند: *زیر پیراهنی، پیش پرداخت، پس لرزه، پس انداز، پیش آمد، بالا پوش، پایین دست*.

این نوع اسم‌های مرکب نیز طبقهٔ بازی را در فارسی تشکیل داده و در پیکره به عنوان یک اسم در نظر گرفته شده‌اند.

برخی اسم‌های مرکب از صفت‌های مرکب به علاوهٔ پسوند اسم‌ساز (i-) به وجود می‌آیند مانند: از خود گذشتگی، پیش پا افتادگی، دو زبانگی.

این دسته اسم‌های مرکب، برچسب‌های اجزای کلام مجزا گرفته‌اند و بنا به دلایلی که برای صفت‌های مرکب سازندهٔ این اسم‌ها در بخش پیشین ذکر شد، بهتر است که این نوع آرایش‌های ترکیبی نیز به عنوان اسم در نظر گرفته شوند.

اسم‌های مرکب همچنین می‌توانند از ترکیب اسم‌هایی چون «میان، پایان» با «اسم» دیگری حاصل شوند مانند: *میان دوره، میان ترم، پایان نامه، پایان خدمت*.

بنابراین اگر یک تحلیل‌گر صرفی خوب وجود داشته باشد، نیازی به افزودن این اسم‌های مرکب در واژگان وجود ندارد و آنها دارای آرایش ترکیبی بوده و طبقهٔ بازی را تشکیل می‌دهند.

۶-نتیجه‌گیری

در این مقاله به بررسی برچسب‌گذاری صرفی و نحوی واحدهای چندقطعه‌ای در متون فارسی به طور پیکره‌بنیاد پرداختیم. به خاطر سرهم بودن نوشتار فارسی و تنوع نوشتاری واژه‌های یکسان [۱۰]،[۱۱]و [۱۸] و نیز به دلیل وجود آرایش‌های ترکیبی و غیرترکیبی قطعه‌ها و امکان ایجاد هر دو آرایش در برخی قطعه‌ها که منجر به برچسب‌های صرفی و نحوی مختلف می‌شوند، و نیز وجود آرایشهای نیمه‌ترکیبی در قطعه‌ها، دو نوع واحد چندقطعه‌ای ایستا یا غیرزایا و پویا یا زایا تعریف گردید و روی مقوله‌های فعل و مصدر، حرف‌اضافه و ربط، قید، صفت و اسم اعمال شد.

واحدهای چندقطعه‌ای ایستا که مربوط به ساختارهای غیرزایای زبان و معمولاً آرایش غیرترکیبی هستند در اکثر موارد طبقه‌های بسته‌ای را تشکیل می‌دهند مانند حروف اضافه و ربط مرکب، قیدهای مرکب و تاحدی افعال مرکب [۱۲]،[۱۹]،[۲۰]و[۲۱]. جهت برچسب‌دهی آرایش‌های ترکیبی که به صورت زایا عمل می‌کنند، در صورت امکان به منظور عدم کاهش زایایی گروه‌های نحوی به مقوله‌های صرفی، الگوهای فعلی، ربطی، قیدی، صفتی و اسمی تعریف می‌شوند که در آنها گروه نحوی در اکثر مواقع آرایش ترکیبی خود را حفظ کرده و کل گروه نیز یک برچسب نقشی معنایی یا نحوی می‌گیرد. بااین‌وجود، در برخی آرایش‌های ترکیبی و نیز نیمه‌ترکیبی کل گروه نحوی یک برچسب صرفی واحد می‌گیرد و طبعاً ساختار نحوی درونیش پنهان می‌شود که البته در صورت نیاز قابل بازیابی می‌باشد. ارائهٔ راهکارهای برخورد با واحدهای چندقطعه‌ای می‌تواند در ایجاد قواعدی به منظور کاربرد در طراحی و پیاده‌سازی تحلیل‌کننده‌های صرفی، تجزیه‌گرهای نحوی و نیز ایجاد بانک درخت گروه‌های نحوی فارسی بر اساس بانک درخت پن تأثیر مستقیم داشته باشد[۲]، [۳]،[۴]،[۵]و [۲۵].

مراجع

[1] W. N Francis H. Kučera "Brown Corpus Manual: Manual of Information to accompany a Standard Corpus of Present-Day Edited American English for use with Digital Computers, Revised and Amplifier Edition". Providence; Dept. Of Linguistics, Brown University. 1979.

[2] M. Marcus, B. Santorini, and M. A. Marcinkiewicz "Building a large annotated corpus of English: The Penn Treebank", Computational Linguistics 19(2), 313-330, 1993.

[3] B. Santorini and M.A. Marcinkiewicz "Bracketing guidelines for the Penn Treebank Project" Ms.,

در این خصوص به مثال‌های (۶) تا (۱۱) توجه نمایید. بند وابسته به صورت پررنگ نشان داده شده و نقش بند وابسته درون پرانتز آورده شده است.

(۶) او روزنامه را **در حالی** می‌خواند **که داشت می‌خوابید**. (ادات فعل)

(۷) **پس از این که درسش تمام شد**، خوابید. (ادات جمله)

(۸) جثه‌اش **قبل از اینکه شروع به ورزش کند**، کوچک بود. (ادات اسم)

(۹) زمان مناسب **بعد از وقتی** است **که کارش تمام می‌شود**. (گزاره)

(۱۰) **با فرض آن که نگار درست می‌گوید**، موضوع را تایید کردیم.(متمم حرف‌اضافه مرکب)

(۱۱) **با در نظر گرفتن این موضوع که نمی‌آیند**، ما رفتیم.(متمم **PP**)

۵-۳- قیدها

در این بخش به بررسی «الگوهای قیدی واحدهای چندقطعه‌ای[۴۴]» و نقشی که برچسب اجزای کلام در کمک به تعیین این برچسب نقش معنایی دارد، می‌پردازیم.

بسیاری از گروه‌های حرف‌اضافه‌ای که واحدهای چندقطعه‌ای پویا هستند می‌توانند نقش قیدی ایفا کنند. یکی از این الگوها ازگروه‌های حرف‌اضافه‌ای شامل دو قطعهٔ حرف‌اضافه و اسم مفرد عام تشکیل شده‌اند مانند «به» و اسم مفرد عامی که از «صفت» به‌علاوهٔ پسوند اسم‌ساز «**i-**» تشکیل شده است مانند *به سختی*، به نرمی. یا حرف‌اضافهٔ «با» و اسم،عام،مفرد مانند با حرارت، با پیروزمندی، با شور و شوق.

اینها درواقع نقش قیدهای حالت توصیف‌کنندهٔ فعل را دارند و در قلاب‌گذاری نحوی به‌صورت گروه حرف‌اضافه‌ای با برچسب نقشی حالت (**MNR-**) برچسب‌گذاری می‌شوند. مانند: **[PP-MNR be saXti], [PP-MNR bA shur o sho:G]**

اما همانگونه که گفته شد اگر این ترکیبات در نوشتار، بدون فاصله و یا با فاصلهٔ مجازی نوشته شده باشند، در قلاب‌گذاری به عنوان گروه قیدی برچسب‌گذاری می‌شوند مانند: **[ADVP-MNR besaXti]**.

همان‌گونه که قبلاً گفته شد اینگونه ترکیبات الزاماً معنای غیرترکیبی و نقش حالت ندارند و در بسیاری موارد صرفاً یک گروه حرف‌اضافه‌ای با معنای ترکیبی هستند. برای نمونه به مثال (۱۲) توجه کنید. در جملهٔ (۱۲-الف) «به سختی»، نقش قید حالت دارد که فعل را توصیف می‌کند اما در جملهٔ (۱۲-ب) چنین نیست.

(۱۲-الف) ساسان به سختی کار می‌کرد.

(۱۲-الف) ساسان به سختی عادت کرده بود.

از اطلاعات درون پیکره و نیز اطلاعات نحوی نمی‌توان به طور دقیقی این نقش‌های معنایی را تعیین کرد. ازاین‌رو، پس از قلاب‌گذاری نحوی، کار اصلاح دستی نیز باید صورت پذیرد.

الگوی دیگر، گروه‌های حرف‌اضافه‌ای با نقش معنایی قیدحالت هستند که از قطعه‌های ***«به (طور/طرز/شکل/گونهٔ/نوع/صورت/شیوهٔ)»*** و یک ***«گروه صفتی»*** تشکیل شده‌اند مانند: *به طور وحشتناکی، به شکل خارق‌العاده‌ای، به شیوهٔ نسبتاً عامیانه.*

این موارد نیز اگر به صورت بدون فاصله یا با فاصلهٔ مجازی در نوشتار ظاهر شوند، برچسب نحوی گروه قیدی می‌گیرند. در غیر اینصورت، به عنوان گروه حرف‌اضافه‌ای در نظر گرفته شده و برچسب نقشی معنایی حالت (**MNR-**) دریافت می‌نمایند.

دسته‌ای از گروه‌های قیدی حالت، تشکیل شده از اعداد اصلی و کلمات دارای برچسب متفرقه،صفت‌ساز هستند مانند *دو نفره، سه تایی*. همانگونه که در بخش ۳-۲ عنوان شد، این آرایش‌های نیمه‌ترکیبی را می‌توان به صورت (**ADVP-MNR**) برچسب‌گذاری کرد. البته این ترکیبات به عنوان صفت مرکب نیز می‌توانند به کار روند که در بخش بعد به آن اشاره می‌شود.

الگوی دیگر شامل مواردی چون *به عنوان مثال، به طور نمونه*، حاوی اسم‌هایی است که نشاندهندهٔ مثال هستند. در پیکره کلماتی چون «مثلاً» برچسب **قید،کلی،مثال** گرفته‌اند. بنابراین پیشنهاد می‌شود که گروه‌های حرف‌اضافه‌ای مانند **به عنوان مثال**، به صورت یک واحد چندقطعه‌ای در نظر گرفته شده و برچسب نحوی گروه قیدی بگیرند. البته در بانک درخت پن، برچسب نقشی معنایی برای «مثال» در نظر گرفته نشده است که در صورت نیاز می‌توان چنین برچسبی را برای فارسی تعریف کرد(مثلاً **EXM-**).

یکی دیگر از الگوهای قیدی واحدهای چندقطعه‌ای، گروه‌های اسمی هستند که می‌توانند نقش معنایی مکان (**LOC-**) داشته باشند و از توصیف‌گرهای اشاره مانند *این(اون)، آن(اون)، همین، همان*، و اسم‌های عام مانند *جا، بالا، پایین، زیر*، تشکیل شده‌اند. کلماتی مانند *اینجا، آنجا* بسته به نقش دستوریشان، دو نوع برچسب *قید،کلی،مکان* و *اسم،عام،مفرد* در پیکره گرفته‌اند البته فراوانی برچسب قید بسیار بیشتر است(مثلاً ***اینجا***(قید،کلی،مکان: ۱۲۵۶و اسم،عام،مفرد:۷۸)).

هنگام قلاب‌گذاری، این گونه گروه اسمی به صورت (**NP-LOC**) برچسب‌گذاری می‌شوند.

۵-۴- صفت‌ها

برخی واحدهای چند قطعه‌ای چون «اسپانیایی زبان»، در پیکره برچسب مجزا گرفته‌اند و همانگونه که پیشتر در بخش ۳-۲ ذکر شد شاید به این دلیل بوده که علاوه بر آرایش غیرترکیبی، آرایش ترکیبی نیز برای این قطعه‌ها در پیکره دیده می‌شود.

یکی از «الگوهای صفتی واحدهای چندقطعه‌ای[۴۵]» مربوط به ترکیب حروف‌اضافه در نقش پیشوند و سایر اجزاء است. یک گروه صفت‌های چندقطعه‌ای زایایی هستند که از ترکیب پیشوند ***«از (پیش/قبل)»***، **«اسم»** و فعل،صفت‌مفعولی **«شده»** به دست می‌آیند مانند: *از پیش تعیین شده، از قبل عنوان شده.*

پیشوند دیگری که به همراه یک صفت دیگر، یک صفت مرکب ایجاد می‌کند، پیشوند **«از خود»** است. مانند*از خود بیگانه، از خود رها.*

این قطعه‌ها هنگام سرهم نوشته شدن، برچسب واحد «صفت» می‌گیرند ولی در حالت جداازهم برچسب‌های مجزا در پیکره دریافت نموده‌اند.

در گروه بندی نحوی چنین صفت‌های مرکبی ممکن است بگوییم که این‌ها گروه‌های حرف‌اضافه‌ای هستند که نقش توصیفی دارند (**PP-ADJ**). اما از آنجاییکه در برچسب‌گذاری اجزای کلام، اگر این قطعه‌ها به صورت بدون فاصله یا با فاصلهٔ مجازی نوشته شده باشند، برچسب «صفت،ساده» می‌گیرند و از سویی دیگر، این گروه‌های حرف‌اضافه‌ای که به عنوان پیشوند هستند در زبان انگلیسی دارای معادل‌هایی چون پیشوندهای «**pre-**» و «**self-**» دارند و در واژه‌نامه‌های آن زبان این آرایش‌های ترکیبی به عنوان صفت در نظر گرفته می‌شوند، لذا پیشنهاد می‌شود در گروه‌بندی نحوی، صفت‌های چندقطعه‌ای را به صورت یک واحد صفتی در نظر گرفته و ساختار نحوی درون آن را نیز پنهان نماییم.

اسم‌های چندقطعه‌ای در اکثر موارد به‌درستی به عنوان یک واحد در نظر گرفته شده و برچسب‌گذاری شده‌اند.

دسته‌ای از این اسم‌های مرکب تشکیل شده از دو اسم می‌باشند که اسم اول همیشه بدون کسرهٔ اضافه بوده و اسم دوم ممکن است در نقش «پسوند فاعلی» مانند «دار» باشد. این نوع ترکیبات معمولاً تشکیل طبقهٔ بسته‌ای را می‌دهند و به صورت زایا اسم مرکب تولید نمی‌کنند مانند *سوار کار، داروخانه دار*. قاعدهٔ شناسایی این نوع اسم‌های مرکب آن است که اولاً قطعه‌ها درون یک گروه باشند و ثانیاً قطعهٔ اول کسرهٔ اضافه نگیرد. برخی از این ترکیب‌ها مانند «پایان خدمت» برچسب مجزا گرفته‌اند و دلیل آن نیز احتمالاً وجود آرایش‌های ترکیبی در پیکره است. بااین‌همه، اسم‌های مرکب در زبان فارسی معمولاً طبقهٔ بازی را تشکیل داده و به صورت زایا و پویا عمل می‌کنند. در بخش ۵-۵ به اسم‌های چندقطعه‌ای پویا می‌پردازیم.

۵-واحدهای چندقطعه‌ای پویا

بیشتر قطعه‌هایی که نقش افعال ،حروف اضافه، حروف ربط، قیدها، صفت‌ها و اسم‌های مرکب را ایفا می‌کنند به عنوان یک واژه و با یک برچسب صرفی در نظر گرفته می‌شوند لذا ساخت نحوی درونیشان پنهان می‌شود.

در این بخش به واحدهای چندقطعه‌ای اشاره می‌کنیم که معمولاً طبقهٔ باز و پویایی را تشکیل می‌دهند که الگوهای خاصی نیز برای آنها قابل تعریف می‌باشد.

۵-۱- افعال و مصدرها

همانطور که در بخش ۴-۱ دیدیم، افعال و مصدرهای چندقطعه‌ای، افعال و مصدرهای مرکب هستند که برخی طبقهٔ بسته‌ای را در زبان تشکیل می‌دهند و زایایی زیادی ندارند و برخی دیگر طبقهٔ بازی را ایجاد می‌کنند.

پر بسامدترین فعل سبکی که در پیکره وجود دارد تصریفهای مصدر «کردن» و زمان گذشتهٔ سادهٔ آن دارای فراوانی ۴۴۵۶۰ است. این فعل کاربرد فعل اصلی بودنش را از دست داده است و تقریباً همیشه سازندهٔ یک فعل مرکب است. فعل مرکب «نمودن» نیز با فراوانی ۲۴۲۹ در پیکره، زایایی زیادی در فارسی دارد.

بسیاری از واژه‌ها که در زبان انگلیسی هم به عنوان اسم به کار می‌روند و هم فعل، در فارسی با افزودن فعل سبک «کردن»، «نمودن» و «زدن» به اسم آن فعل مرکب را ایجاد می‌کنند مانند: *تلفن زدن، پست کردن، ایمیل کردن*، فرمت نمودن، بوت کردن، اس‌اس زدن و ... .

در مورد بقیهٔ افعال سبک، برخی زایایی نسبتاً زیادی دارند مانند «آوردن» و «بردن» که طبقهٔ بازی را ایجاد می‌کنند ولی برخی دیگر مانند «شستن» و «چیدن» زایایی کمتری داشته و طبقهٔ بسته‌ای را تشکیل می‌دهند.

جهت افعال مرکب با زایایی زیاد که طبقهٔ بازی را تشکیل می‌دهند می‌توان «الگوهایی فعلی واحدهای چندقطعه‌ای» پویا را تعریف کرد.

الگوی اول ترکیب جزءهای غیرفعلی اسمی، صفتی، حرف‌اضافه‌ای با پربسامدترین افعال سبک یعنی «کردن» و «نمودن» است.

الگوهای دیگر را نیز می‌توان برای سایر افعال سبک تعریف کرد. نکته‌ای در تعریف این الگوها باید توجه شود آن است که برای هریک از افعال سبک موجود در بانک اطلاعاتی واژگان، باید اجزای غیرفعلی از پیکره استخراج شده و در جدول‌هایی که به نام اجزای غیرفعلی هستند درج شوند. این جدول‌ها به صورت یک‌به‌چند، با جدول افعال سبک مرتبط شده‌اند. مثلاً در مورد فعل «آوردن» می‌توان الگوی (۴) را که از پیکره استخراج شده است ارائه کرد:

(۴)

[vP [PredP x] [v 'Avardan]
x ε {N, Adj, Particle, PP}

N={'Ab, bAr, jA, hojum, forud, padid, farAham, ra'y, tAb, davAm, vAred, feshAr,'imAn,'ozr, bahAne, ruy, tashrif, rahm}

Adj={gerd, kam}

Particle={bAz, bar, dar, bar, birun, pAyin, pish, foru}

PP={ be jA(y), be chang, be hesAb, be XAter, be Xashm, be Xod, be dar, be dard, be dast, be ruy-e kAr, be zabAn, be shomAr, be shur, be sahne, be'amal, be kaf, be miyAn, be nazar, be vajd, be vojud, be hamrAh, be hush, be yAd}

ساخت‌های مجهولی که با فعل کمکی «شدن» ایجاد می‌شوند نیز یکی دیگر از الگوهای فعلی واحدهای چندقطعه‌ای پویا را به وجود می‌آورند. این نوع ساخت‌ها در فارسی می‌توانند به عنوان فعل مرکب در نظر گرفته شوند [۲۰]. بنابراین صفت مفعولی ایجادشده از تمام افعال متعدی به همراه فعل «شدن»، یک الگوی فعلی چندقطعه‌ای پویا ایجاد می‌نماید مانند: *دیده شده اند، از میان برده شد*.

۵-۲- حروف اضافه و ربط

قطعه‌های چون *(قبل/پیش) از* و *(بعد/پس) از* را نمی‌توان یک واحد در نظر گرفت زیراکه امکان انجام همپایگی درون اجزای آنها وجود دارد مانند: *قبل و بعد از ظهر، پیش و پس از انقلاب*.

برخلاف بانک درخت پن که کلاً حروف‌اضافهٔ چند واژه‌ای زبان انگلیسی (مانند **because of**) را به صورت ساختار مسطح و یک حرف‌اضافه در نظر می‌گیرد، بهتر است این‌گونه موارد را به عنوان دو گروه حرف‌اضافه‌ای درون هم در نظر بگیریم تا بتوانیم ساخت‌های همپایه را توجیه کنیم مثلاً *قبل از انقلاب* به صورت (۵) گروه‌بندی نحوی کنیم.

[PP-TMP Gabl [PP 'az [NP 'enGelAb]]] (۵)

با توجه به آنکه حروف اضافهٔ چندقطعه‌ای از نوع مرکب و ایستا هستند و طبعاً طبقهٔ بسته‌ای را تشکیل می‌دهند، لذا «الگوی حرف‌اضافهٔ چندقطعه‌ای»[۴۲] مفید و موثری برای آن نمی‌توان تعریف کرد.

در بخش ۴-۲ به ساختار نحوی حروف ربط مرکب که دراصل واحدهای چندقطعه‌ای ایستا هستند اشاره نمودیم. در این بخش، به بررسی حروف ربط چندقطعه‌ای و «الگوی ربطی واحدهای چندقطعه‌ای[۴۳]» می‌پردازیم.

آن دسته از حروف ربط که از حروف‌اضافهٔ مرکب به‌علاوهٔ (اینکه/آنکه) یا گروه حرف‌اضافه‌ای به‌علاوهٔ (که) به وجود آمده‌اند، به صورت متمم حرف‌اضافه برچسب‌گذاری نحوی می‌شوند مثلاً:

به علت اینکه

[PP-PRP be 'ellat-e [NP [NP 'in] [CP ke [vP...]]]]

به این علت که

[PP-PRP be [NP [NP 'in 'ellat] [CP ke [vP...]]]]

و یا

با در نظر گرفتن این فرض که

[PP bA [NP [NP dar nazar gereftan-e 'in farz] [CP ke [vP...]]]]

دلیل، نیز حرف‌اضافهٔ مرکب نیستند بلکه گروه حرف‌اضافه‌ای هستند. نکتهٔ قابل توجه در این ترکیبات آن است که کسرهٔ اضافه ندارند و اگر کسرهٔ اضافه بگیرند بدساخت می‌شوند و در واقع ترکیباتی چون *بر این اساس* و *به این دو دلیل*، گروه حرف‌اضافه‌ای هستند که نقش حروف ربط مرکب را ایفا می‌کنند.

حروف اضافه در زبان‌ها معمولاً یک طبقه بسته را تشکیل می‌دهند. وجود این ترکیبات حرف‌اضافه و اسم با توجه به شرایط بالا، این فرض را حمایت می‌کند که در زبان فارسی نیز حروف اضافهٔ مرکب یک طبقهٔ بسته را تشکیل بدهند. با توجه به بررسی‌های انجام‌شده در پیکره، ۳۸ حرف‌اضافهٔ مرکب یافت شد که در حالت سرهم برچسب حرف‌اضافه گرفته‌اند. اما با توجه به دو شرط فوق و بررسی‌ها و تست‌های صرفی و نحوی انجام شده، حدود ۷۰ حرف‌اضافهٔ مرکب در فارسی وجود دارد. مثلاً قطعه‌هایی چون « به خاطرِ/دلیلِ/سببِ» حرف‌اضافهٔ مرکب هستند اما در پیکره حتی هنگامی‌که سرهم نوشته شده‌اند به عنوان قطعه‌های چندواحدی در نظر گرفته شده و برچسب «حرف‌اضافه،اسم،مفرد،عام، کسره‌اضافه» گرفته‌اند که این مساله باید در پیکره اصلاح شود. از این ۱۰۵ حرف‌اضافه، حدود ۴۰ حرف می‌توانند با «(این/آن) که» ترکیب شده و گروه حرف‌اضافه‌ای با متمم‌های گروه متممی و نقش‌های معنایی همچون علت (**PRP-**) به وجود آورند. در این حالت آنها واحدهای چندقطعه‌ای پویا به وجود می‌آورند که در بخش بعدی بحث می‌شوند.

سایر حروف ربط مرکب از جمله *اگر چه*، *با این وصف*، *هر طور که*، *به هیچ وجه*، *از هنگامی که*، *که به موجب آن*، *در صورتی که*، *در تمام مدتی که*، *به قدری که*، نیز حدود ۱۳۰ حرف ربط مرکب را ایجاد می‌کنند و در مجموع تعداد حروف ربط چندقطعه‌ای زبان فارسی حدود ۲۱۰ حرف می‌باشد.

درواقع، حروف ربط چندقطعه‌ای در زبان فارسی، حروف ربط وابستگی هستند. توزیع حروف ربط مرکب و چندقطعه‌ای درپیکره شامل ادات جمله‌ای/فعلی، ادات یا متمم اسم، گزاره، متمم گروه فعلی و متمم حرف‌اضافه مرکب و یا گروه‌حرف‌اضافه‌ای است.

در بانک درخت پن انگلیسی برای حروف ربط چندواژه‌ای (مانند **as though, in order to, whether or not**) یک ساختار مسطح در نظر گرفته‌اند و برای دستهٔ دیگری از آنها مانند **if not, rather than** ساختار گروه حرف‌ربط(**CONJP**) را اتخاذ نموده‌اند . در برچسب‌گذاری نحوی یک حروف‌ربط چندقطعه‌ای در زبان فارسی ما دو ساختار ایستا و پویا را پیشنهاد می‌کنیم که در این بخش به ساختار ایستا و در بخش ۵-۲ به ساختار پویا اشاره می‌کنیم.

در ساختار ایستا حروف ربط مرکب به عنوان یک واحد در نظر گرفته شده و مانند حرف‌ربط «که» درون هستهٔ یک گروه متممی قرار گرفته و برچسب‌گذاری می‌شوند مانند: با این وجود، اگر چه، با این همه، همان گونه که، گذشته از این.

[CP (bA 'in vojud) (agar che) (bA 'in hame) (hamAn gune ke) (gozashte 'az 'in) [vP…]]

در بخش ۵-۲ به حروف ربط چندقطعه‌ای پویا می‌پردازیم.

۴-۳- قیدهای مرکب

همانند حروف‌ربط و اضافهٔ مرکب که طبقهٔ بسته‌ای را تشکیل می‌دهند، قید های مرکب چندقطعه‌ای نیز وجود دارند که معمولاً از ترکیب یک «حرف‌اضافه» و یک یا دو مقولهٔ «اسم» به وجود می‌آیند.

نکتهٔ مهمی که در این‌باره وجود دارد آن است که اگر یک ترکیب «حرف‌اضافه» و مقولهٔ دیگر مثلاً «اسم» حالت زایا داشته باشد، بهتر است که آن ترکیب را «قید مرکب» نگوییم زیرا در آنصورت یک ساختار پویا و زایای نحوی را به یک مقولهٔ صرفی کاهش داده‌ایم و از سویی طبقهٔ بسته نیز نمی‌باشد. اما اگر به لحاظ نوشتاری قطعه‌ها به صورت بدون فاصله یا با فاصلهٔ مجازی نوشته شوند، از آنجاییکه در پیکره برچسب واحد قید گرفته‌اند مانند **بسختی**، در برچسب‌گذاری نحوی نیز بهتر است به صورت گروه قیدی قلاب‌گذاری می‌شوند.

اکنون به بررسی انواع قیدهای مرکب می‌پردازیم. این نوع قیدها از ترکیب غیرزایای برخی حروف اضافه با مقوله‌های دیگر حاصل می‌شوند. این قیدهای مرکب دارای فراوانی قابل‌ملاحظه‌ای نیز در متون فارسی می‌باشند مثلاً «درواقع» فراوانی۱۴۷۲ و «به تدریج» فراوانی ۴۶۹ دارد. تعداد ۹۰ قید مرکب با نوشتار سرهم در پیکره یافت شد، از جمله: *به اجبار*، *به اختصار*، *به اشتباه*، *به تدریج*، *به تناوب*، *به جرات*، *به خوبی*، *به درستی*، *به دقت*، *به خصوص*، *به سرعت*، *به مراتب*، *به مرور*، *بی پروا*، *بی اختیار*، *بی درنگ*، *بی قید و شرط*، *بی مقدمه*، *در ابتدا*، *در اصل*، *در مجموع*، *در نهایت*، *در واقع*.

از آنجاییکه بسیاری از گروه‌های حرف‌اضافه‌ای می‌توانند نقش قیدی ایفا کرده و به صورت پویا عمل کنند، در بخش ۵-۳ به آن می‌پردازیم.

۴-۴- صفت‌های مرکب

در این بخش به گروه دیگری از واژه‌های چندقطعه‌ای یعنی صفت‌های مرکب می‌پردازیم.

دسته‌ای از صفت‌های چندقطعه‌ای به صورت دو قطعهٔ صفت و اسم (غیرمصدری) هستند مانند: *سیاه بخت*، *گران قیمت*.

پیشوندهایی شامل «**میان، زیر، پایین، بالا، پیش، پس**» به همراه صفت‌های دیگر می‌توانند صفت‌های مرکبی را ایجاد کنند. مانند *میان دوره‌ای*، *زیر آبی*، *پیش گزیده*، *بالا/ پایین/پیش/پس رونده*. همچنین ترکیب پیشوند «شبه» با صفت‌های دیگر، سازندهٔ صفت‌های مرکب است مانند: *شبه دولتی*، *شبه سینمایی*، *شبه غربی*. دستهٔ دیگر صفت‌ها به همراه «فعل،صفت‌مفعولی» هستند مانند *گرما زده*، *تب آلوده*.

این نوع واحدهای چندقطعه‌ای در پیکره برچسب واحد گرفته و به عنوان صفت مرکب در نظر گرفته می‌شوند.

حرف‌اضافهٔ «**به**» و برخی اسم‌ها می‌توانند تولیدکنندهٔ صفت مرکب باشند که در پیکره برچسب مجزا گرفته‌اند ولی طبقهٔ بسته‌ای را تشکیل می‌دهند مانند: *به جا*، *به خصوص*، *به قاعده*.

با بررسی‌های انجام‌شده بر روی پیکره، آن دسته از قطعه‌ها که تشکیل طبقهٔ بسته‌ای از صفت‌ها را داده و عمدتاً در آرایش‌های غیرترکیبی به کار می‌روند، به عنوان صفت مرکب در نظر گرفته شده و طبعاً برچسب واحدی نیز در سطح اجزای کلام دریافت نموده‌اند.

بسیاری از صفت‌ها به صورت زایا از ترکیب قطعه‌های پیشوندی «پر»، «با» و «بی» با اسم حاصل می‌آیند مانند «با حوصله»، «پر کار» و «بی دقت». ازآنجاییکه این پیشوندها به تنهایی ظاهر نمی‌شوند، این‌گونه ترکیبات در پیکره با مقولهٔ صفت در سطح اجزای کلام برچسب‌گذاری شده‌اند.

در بخش ۵-۴ به بررسی صفت‌های چندقطعه‌ای پویا خواهیم پرداخت.

۴-۵- اسم‌های مرکب

گروه دیگری از واژه‌های چندقطعه‌ای، اسم‌های مرکب هستند. درپیکره،

حرف‌اضافه‌ای(*به دنیا آمدن*)، ادات(*کنار کشیدن*)، صفت(*کوتاه کردن*)، اسم(*دست زدن*) و نیز اسم گزاره‌ای[۳۹](*انجام دادن، شکست خوردن*) و صفت گزاره‌ای[۴۰](*فراموش کردن، گم شدن*)باشد(جهت اطلاعات بیشتر در مورد اسم و صفت گزاره‌ای، مراجعه کنید به [۲۳]).

پر بسامدترین فعل سبکی که در پیکره وجود دارد تصریف‌های فعل «کردن» است که زمان گذشتهٔ سادهٔ آن دارای فراوانی ۴۴۵۶۰ و زمان حال ساده‌اش دارای فراوانی۲۶۵۹۴ است. این فعل، دیگر فعل اصلی نبوده و تقریباً همیشه سازندهٔ یک فعل مرکب است. از جمله افعال سبک دیگر می‌توان به *شدن، خوردن، زدن، دادن، داشتن، آوردن، بردن، بودن، پاشیدن، آمدن، گذاشتن، شستن، سپردن، بستن، انداختن، افتادن، کشیدن، گرفتن، چیدن، رفتن و نمودن* اشاره کرد[۲۴]. فعل «*شدن*» در ساخت‌های مجهول و غیرمفعولی به کار می‌رود. سایر افعال سبک در کاربرد فعل اصلی نیز می‌توانند ظاهر شوند. با توجه به اینکه عناصر **NV** و **LV** یک فعل مرکب در سطح نحو جدای از هم هستند، فعل سبک به عنوان هستهٔ یک گروه فعلی(**vP**) بوده و جزء غیرفعلی فعل مرکب درون یک گروه اسناد (**PredP**) قرار می‌گیرد.

مصدرهای مرکب مانند *نگاه کردن* و *دست شستن نیز* به صورت دو عنصر همانند فعل مرکب برچسب‌گذاری شده‌اند. فراوانی مصدر «کردن» در پیکره ۶۲۲۹ است. عنصر فعل سبک، چون دارای وند مصدری است، به صورت «اسم،عام،(مفرد/جمع)،مصدر» برچسب‌گذاری شده است. ازسویی، عناصر مصدرهای مرکب همانند افعال مرکب، معمولاً بدون کسرهٔ اضافه هستند و از سویی دیگر، معمولاً جزء غیرفعلی بدون قرارگرفتن عنصری بعد از آن قبل از فعل سبک قرار می‌گیرد زیرا افعال کمکی چون *خواستن* و *داشتن* که در افعال مرکب می‌تواند آورده شود، در مصدرهای مرکب ظاهر نمی‌شود. به این دلیل، تشخیص مصدر مرکب راحت‌تر می‌گردد.

در بانک درخت پِن انگلیسی جهت گروه بندی نحوی مصدرها آن را به عنوان گروه فعلی با فاعل تهی در نظر گرفته و به گروه فعلی یک برچسب نقشی نحوی «اسمی» (NOM-) می‌دهند [۲۵]. دلیل این کار آن است که ارتباط نزدیک میان ساخت مصدری و فعلی در انگلیسی وجود دارد. در زبان فارسی وجود مصدرها در جمله‌ها، رفتار اسمی قوی‌تری از خود نشان می‌دهند تا رفتار فعلی. بهمین جهت در گروه‌بندی نحوی، پیشنهاد می‌شود که مصدرهای مرکب همانند مصدرهای ساده به صورت گروه اسمی قلاب‌گذاری شوند. دراین‌خصوص به مثال (۳) که قلاب‌گذاری پیشنهادی یک جمله حاوی مصدرهای مرکب در فارسی است توجه کنید.

در گروه‌بندی نحوی، پیشنهاد می‌شود که مصدرهای مرکب همانند مصدرهای ساده به عنوان گروه فعلی با فاعل تهی در نظر گرفته شوند و به گروه فعلی یک برچسب نقشی نحوی «اسمی»(**NOM-**) داده شود. در بانک درخت پِن انگلیسی نیز چنین عمل شده است [۲۵]. دراین‌خصوص به مثال (۳) که قلاب‌گذاری پیشنهادی یک جمله حاوی مصدرهای مرکب است توجه کنید.

[vP [NP-SBJ [NP jang kardan] [PP bA to]] [PredP [NP-PRD [NP jang kardan] [PP bA man]] 'ast]] (۳)

اگرچه بیشتر افعال مرکب زایایی داشته و طبقهٔ بازی را تشکیل می‌دهند، می‌توان «الگوهایی فعلی واحدهای چندقطعه‌ای»[۴۱] ایستا را برای افعال با بسامد کمتر مانند «چیدن» و «شستن» تعریف کرد، اگرچه چنین کاری توصیه نمی‌شود.

۴-۲- حروف اضافه و ربط مرکب

شاید به جرأت بتوان گفت که بیشترین فراوانی کلمات چندقطعه‌ای در متون فارسی مربوط به حروف ربط و حروف اضافهٔ مرکب است. همانطورکه در بخش ۳-۲ اشاره شد، قطعه‌های این مقوله‌های خاص نیز در پیکره برچسب‌های مجزا گرفته‌اند.

در زبان‌هایی مانند انگلیسی برخی از واژه‌ها می‌توانند هم نقش حروف اضافه و هم حروف ربط وابسته را ایفا کنند مانند **after, before**. یک واژه، به عنوان حرف ربط وابستگی تلقی می‌شود اگر پس ازآن یک جمله بیاید و به عنوان یک حرف اضافه است وقتیکه بعد از آن گروه‌های اسمی و یا دیگر متمم‌ها ظاهر شوند[۲۵].

در زبان فارسی برخی حروف اضافه و حروف ربط مرکب شبیه به هم می‌باشند ولی در انتهای حروف ربط مرکب، واژه‌های *اینکه* یا *آنکه* ظاهر می‌شود مانند: *به (خاطر/سبب) (این/آن) که* و *به مجرد (این/آن) که*.

در این گونه موارد درواقع، اجزاء قبل از «*که*» به عنوان گروه حرف‌اضافه‌ای با هستهٔ مرکب بوده و خود «که» و جملهٔ بعد از آن به عنوان متمم در نظر گرفته می‌شوند. در تولید بانک درخت فارسی، جهت شناسایی دقیق بندهای وابسته لازم است تا حروف‌ربط و اضافهٔ مرکب به‌درستی شناسایی شوند.

حروف‌اضافهٔ ساده‌ای که می‌توانند به انضمام یک اسم دیگر یک حرف‌اضافهٔ مرکب ایجاد کنند شامل « *در، به، با، از، بر*» هستند. این حروف کسرهٔ اضافه نمی‌گیرند و مشخصه‌های **[-V, -N]** دارند و به عنوان حروف‌اضافهٔ «واقعی» هستند [۲۶].

در مورد این‌که حروف‌اضافه مرکب کدام‌ها هستند اتفاق نظر وجود ندارد. حروف‌اضافهٔ مرکب که از یک حرف‌اضافه و یک اسم تشکیل شده‌اند، ظاهراً شبیه به یک گروه حرف‌اضافه‌ای هستند. اما شواهد ساختواژی و نحوی وجود دارد که نشان می‌دهد این دو با هم متفاوت هستند، مثلاً امکان درج توصیف‌گر اشاره بین حرف‌اضافه و اسم در یک گروه حرف‌اضافه‌ای وجود دارد(مانند *در **این** کارِ خیر*) اما در یک حرف اضافه مرکب چنین کاری عملی نیست (مانند **به **این** دلیلِ مطالعات*)(جهت ملاحظهٔ شواهد صرفی و نحوی دیگر مراجعه کنید به [۲۷]).

به نظر نگارنده، دسته‌ای از حروف اضافهٔ مرکب شامل قطعه‌هایی هستند که هریک به تنهایی برچسب «حرف‌اضافه» خورده باشند و نیز همواره کنار هم بیایند مانند: *از برای، تا به*. دستهٔ دیگر از یک حرف‌اضافه ساده و یک یا دو اسم با شرایط زیر تشکیل شده باشند:

۱- دارای فراوانی زیادی در زبان بوده و به صورت نوشتاری چسبیده به هم در پیکره یافت شوند.

۲- به لحاظ رفتار صرفی و نحوی رفتاری متفاوت با گروه‌های حرف‌اضافه‌ای شامل حرف‌اضافه و اسم داشته باشند.

با توجه به شرایط بالا قطعه‌ها *رو به روی، سر تا سرِ، به وسیلهٔ، در بارهٔ، به دستِ، به علتِ، نزدیک به، عطف به، در پیِ، به جز، به مجرد، به محضِ، به موجبِ*، به علتِ، به دلیلِ، به خاطرِ، حرف‌اضافهٔ مرکب هستند.

این‌گونه حروف اضافهٔ مرکب به صورت یک حرف اضافه(**P**) درنظر گرفته می‌شوند و هستهٔ یک گروه حرف‌اضافه‌ای را تشکیل می‌دهند.

در اینجا ذکر یک نکته لازم است و آن اینکه برخی گروه حرف‌اضافه‌ای وجود دارند که بسیار شبیه به حروف‌اضافهٔ مرکب می‌باشند مانند ***براین اساس، به این دو دلیل***. ممکن است گفته شود که چون بین حرف‌اضافه و اسم، اجزای دیگر ظاهر شده، پس ترکیبات *بر اساس* و *به*

نقش معنایی ممکن است در برچسب اجزای کلام برخی قطعه‌های درون گروه نحوی وجود داشته باشد. مانند « به اشتباه» که یک گروه حرف‌اضافه‌ای است تشکیل شده از یک «حرف‌اضافه» و یک «اسم،عام،مفرد». در جمله‌ای چون «ساسان به اشتباه خو گرفته بود »، معنای ترکیبی دارد اما این گروه نحوی در جملهٔ«ساسان به اشتباه کتاب مرا برداشت» معنای غیرترکیبی و نقش قید حالت (MNR-) دارد.

حالت چهارم که حالت جدیدی می‌باشد و ما آن را «آرایش نیمه‌ترکیبی[۲۹]» می‌نامیم آن است که قطعه‌های درون یک گروه بتوانند با الگوهای خاصی و البته نه به طور کاملاً آزادانه با یکدیگر ترکیب شوند و معنایی ترکیبی نیز حاصل کنند که از این نظر شبیه به آرایش ترکیبی هستند. ولی کل گروه نحوی نیز یک برچسب می‌گیرد که از این حیث شبیه آرایش غیرترکیبی است مانند «۱۳۰۰ کیلویی» و «چهار لیتری» که این گروه‌ها در نقش صفت هستند و «دو نفره» که می‌تواند هم گروه صفتی و هم گروه قیدی باشد.

از چهار حالت مذکور، حالت اول تشکیل‌دهندهٔ یک طبقهٔ باز و زایا در زبان است که قطعه‌ها می‌توانند آزادانه با یکدیگر ترکیب شوند و به عنوان واحد چندقطعه‌ای تلقی نمی‌شوند. حالتهای دوم و سوم و چهارم که مرتبط با معنای غیرترکیبی هستند، در کار با واحدهای چندقطعه‌ای دخیل می‌باشند. حالت دوم یعنی آرایش‌های غیرترکیبی طبقهٔ بسته‌ای را در زبان فارسی تشکیل می‌دهند و زایایی ندارند و به همین دلیل است که در واژگان مدخل‌های جداگانه‌ای دارند و به عنوان واژه‌های مرکب[۳۰] در نظر گرفته می‌شوند. این گونه موارد، تشکیل‌دهندهٔ «*واحد چندقطعه‌ای ایستا*» می‌باشند که در آن کل گروه نحوی به عنوان یک واحد زبانی و با یک برچسب صرفی یا نحوی در نظر گرفته می‌شود و طبعاً ساختار نحوی درون آن نیز ممکن است پنهان شود مانند حروف ربط و قیدهای مرکب ولی در افعال مرکب چنین نیست (بخش ۴-۱).

حالت سوم چون مربوط به گروه‌هایی است که دو آرایش ترکیبی و غیرترکیبی می‌توانند داشته باشند لذا مشکل‌سازترین حالت در برچسب‌گذاری متون است و در برچسب‌گذاری پیکره، این حالت به عنوان واحد چندقطعه‌ای در نظر گرفته نشده است و هریک از قطعه‌ها برچسب مجزا گرفته‌اند و به عبارتی دیگر آرایش ترکیبی برای آنها لحاظ شده است. در این حالت فراوانی آرایش ترکیبی نسبت به غیرترکیبی بیشتر است. در گروه‌بندی نحوی بخش غیرترکیبی آنها، می‌توان واحدهای چندقطعه‌ای را همان گروه نحوی حالت ترکیبی در نظر گرفت و علاوه بر آن یک نقش معنایی نیز به کل گروه نحوی اختصاص داد. مثلاً «به سختی» در جملهٔ «ساسان به سختی کار می‌کرد»، یا «به اشتباه» در جملهٔ «نگار به اشتباه رفتار کرد» همان گروه حرف اضافه‌ای در نظر گرفته شود ولی برچسب نقشی معنای قیدحالت نیز بگیرد و به صورت [PP-MNR] قلاب‌گذاری شود. این حالت طبقهٔ بازی را تشکیل داده و الگو[۳۱]های خاصی نیز برای آن قابل تعریف است و به این دلیل تشکیل‌دهندهٔ «*واحد چندقطعه‌ای پویا*» می‌باشد. در برخی موارد که آرایش غیرترکیبی نسبت به ترکیبی فراوانی بیشتری دارد، علیرغم زایایی ساختار و پویا بودن چندقطعه‌ای، کل گروه نحوی یک برچسب صرفی می‌گیرد و درنتیجه ساختار نحوی درونیش، پنهان می‌شود.

در مورد برچسب‌گذاری حالت چهارم، که از یک سو ترکیبی و از سویی دیگر، غیرترکیبی است، دو گزینه پیش روی ماست. اولین گزینه آنکه گروه نحوی به عنوان یک واژه تلقی شده و یک برچسب صرفی بگیرد که درآن‌صورت ساختار نحوی آن پنهان می‌شود مثلاً «دو نفره» به عنوان یک واژه و با برچسب اجزای کلام صفت و یا قید در نظر گرفته شود.

گزینهٔ دوم آن است که قطعه‌های عددی، به صورت یک گروه نحوی کمیت‌نما در نظرگرفته شده و به همراه قطعهٔ دارای برچسب «متفرقهٔ صفت‌ساز» مانند «نفره، کیلویی، لیتری» درون یک گروه صفتی یا قیدی قرار بگیرد. گزینهٔ اول مناسب نبوده و بر خلاف جهت زایایی زبان است زیرا یک گروه نحوی زایا را به یک مقولهٔ صرفی کاهش داده و ساختار نحوی را پنهان نموده‌ایم. گزینهٔ دوم مناسب‌تر بوده و تشکیل‌دهندهٔ واحد چندقطعه‌ای پویا است.

همانگونه که در ابتدای مقدمه نیز گفته شد، در تولید بانک درخت فارسی به مسألهٔ واحدهای چندقطعه‌ای باید توجه ویژه‌ای شود و یک نظام پردازشگر فرعی طراحی گردد تا کار تقطیع متن فارسی را انجام دهد.

همانگونه که می‌بینیم عدم امکان تقطیع صحیح واحدهای چندقطعه‌ای باعث می‌شود که تحلیل‌کننده‌های صرفی و تجزیه‌گرهای نحوی با مشکلات جدی مواجه شوند.

با توجه به مطالب فوق در دو بخش بعدی با جزئیات بیشتر به واحدهای چندقطعه‌ای ایستا و پویا می‌پردازیم.

۴-واحدهای چندقطعه‌ای ایستا

در بخش قبل اشاره شد که بخشی از واحدهای چندقطعه‌ای، طبقهٔ بسته و غیرزایایی را در زبان تشکیل می‌دهند و همچنین به‌دلیل آرایش غیرترکیبی‌شان، مدخل‌های جداگانه‌ای در واژگان با عنوان واژه‌های مرکب دارند. همانگونه‌که گفته شد این حالت، ایجادکنندهٔ واحدهای قطعه‌ای ایستا است. در زیربخش‌های بعدی به واحدهای چندقطعه‌ای ایستا شامل افعال و مصدرهای مرکب، حروف ربط و اضافهٔ مرکب و نیز قیدها، صفت‌ها و اسم‌های مرکب می‌پردازیم.

۴-۱- افعال و مصدرهای مرکب

فعل مرکب فارسی[۳۲] را نمی‌توان یک واحد واژگانی دانست زیراکه بین عنصر غیرفعلی[۳۳] و فعل سبک[۳۴] ممکن است تعدادی عنصر دیگر ظاهر شوند از جمله فعل کمکی زمان آینده(خواستن) و نیز وندهای نفی و تصریفی[۱۹].

زبانشناسان و مهندسان زبان در تجزیه و تحلیل افعال مرکب فارسی در چارچوب نظریه‌های واژگانی با مشکل جدی مواجه می‌شوند و معمولاً به این نتیجه می رسند که این افعال نمونه‌هایی از «اصطلاحات[۳۵]» هستند که در واژگان باید مدخل جداگانه‌ای به همراه ساختار نحوی‌شان داشته باشند[۲۰]. در [۲۱] بحث شده است که خصوصیات متضاد افعال مرکب فارسی در نظریه‌های غیرواژگانی چون ساختواژهٔ توزیع‌شده[۳۶] به راحتی قابل توجیه هستند که در آن تمام تفسیرها به صورت فرانحوی انجام می‌شوند و بر اساس ساختار پیشنهادی [۲۲]، ساخت نحوی افعال مرکب برای زبان فارسی مشابه‌سازی شده است. مثلاً ساخت نحوی افعال مرکب آغازی[۳۷] مانند *بیدارشدن*، و ساخت سببی آن *بیدارکردن*، به نقل از [۲۰] در (۱) و (۲) نشان داده شده است.

[vP [AdjP [DP Kimea] [Adj bidAr]] [v shod]] (۱)

[vP [DP Parviz] [v' [AP [D Kimea-ro][Adj bidAr]] [v kard]]] (۲)

این مثال‌ها نشان می‌دهند که یک فعل سبک خاص در افعال مرکب فارسی، منجر به ظهور و یا عدم ظهور یک عنصر نقش‌پذیر عامل می‌گردد و البته در نظریهٔ گروه فعلی پوسته‌ای[۳۸] انتظار آن می‌رود.

در برچسب‌گذاری اجزای کلام پیکره، جزء غیرفعلی و فعل سبک افعال مرکب چنین برچسبی نگرفته‌اند. جزء غیرفعلی فعل مرکب می‌تواند گروه

غ، ف، ق، ک، گ، ل، م، ن، ه، ی. بقیه حروف که حروف مجزا خوانده می‌شوند، فقط یک شکل دارند ؛ ا، د، ذ، ر، ز، ژ، و.

همچنین، نظام خط فارسی چنین اجازه می‌دهد که برخی تکواژها یا به صورت وندهای مقید و یا آزاد، قبل و یا بعد از یک تکواژ ظاهر شوند[۱۲]. بنابراین، بسیاری از واژه‌ها به دو صورت سرهم یا جداازهم می‌توانند نوشته شوند به‌طوری‌که واحدهای زبانی سرهم یا جداازهم، تکواژهای سازندهٔ واژه‌ها هستند. در صورت سرهم، شکل آغازی یا میانی تکواژ اول به شکل میانی یا پایانی تکواژ دوم متصل می‌شود مثلاً در واژهٔ «سیبها»، صورت آغازی « ب » از تکواژ «سیب» به صورت میانی « ـبـ » از تکواژ «ها»، متصل شده است. در صورت جداازهم، بین تکواژها، «فاصله»[۲۱] و یا «فاصلهٔ مجازی(نیم‌فاصله)»[۲۲] درج می‌گردد و صورت پایانی تکواژ اول به صورت آغازی تکواژ دوم متصل می‌شود مانند «سیب ها» و «سیب‌ها».

در مورد تعریف واژه، به‌عنوان یک واحد زبانی، میان زبان‌شناسان اتفاق نظر وجود ندارد، اما با دقت قابل‌قبولی می‌توان واژه را در هر زبان تعریف کرد[۱۳]، [۱۴]، [۱۵]. روان‌شناسان زبان در تحقیقات خود ثابت کرده‌اند که، واژه‌ها و تکواژها، واحدهای دستیابی زبان در درک زبان هستند[۱۶]. اگرچه گویشوران زبان و تایپیست‌ها معمولاً به لحاظ شهودی، مشکلی درتشخیص واژه‌ها ندارند، اما به لحاظ نوشتاری به طور یکسانی عمل نکرده و از استاندارد خاصی تبعیت نمی‌کنند. فرهنگستان زبان و ادب فارسی در دستورخط از جدانویسی حمایت کرده است به‌طوریکه بین تکواژهای سازنده، نویسهٔ نیم‌فاصله قرار گیرد[۱۷].

مطالب فوق دربارهٔ نوع حروف و شیوهٔ نوشتار خط عربی و فارسی، علل اصلی در پیدایش آن چیزی است که آن را *تنوع نوشتاری*[۲۳] می‌نامند[۱۸].

فرض کنیم هر *قطعهٔ نوشتاری*[۲۴] توسط یک *مرزنما*[۲۵] که معمولاً نویسهٔ فاصله و یا علائم سجاوندی است، مشخص می‌شود. همچنین این نکته را در نظر بگیریم که هر قطعه می‌تواند یک تکواژ، واژهٔ بسیط، تصریف، اشتقاق و یا ترکیب باشد. درآن‌صورت، یک رابطهٔ چندبه‌چند میان قطعه‌های نوشتاری و واحدهای زبانی به وجود می‌آید. این بدان معنی است که از سویی، یک قطعهٔ نوشتاری ممکن است نشان‌دهندهٔ یک یا چند واحد زبانی باشد مانند قطعهٔ «بمن» که دو واحد زبانی «حرف اضافه» و «ضمیر» است و قطعهٔ «کزین» که می‌تواند سه واحد زبانی «حرف‌ربط»، «حرف‌اضافه» و «ضمیر» باشد. از سویی دیگر، یک واحد زبانی ممکن است توسط یک یا چند قطعهٔ نوشتاری نشان داده شود مثلاً حرف اضافهٔ «به خاطر» که در دو قطعه نوشته شده و دو برچسب صرفی گرفته است و یا حرف‌ربط «با این حال» که در سه قطعه نوشته شده و سه برچسب اجزای کلام نیز در پیکره گرفته است. حالت اول را یک «قطعهٔ چندواحدی(واژه‌ای)»[۲۶] و حالت دوم را یک «واحد(واژهٔ) چندقطعه‌ای» می‌نامیم.

حالت اول یعنی قطعهٔ چندواحدی نیز از اهمیت ویژه‌ای در برچسب‌گذاری نحوی برخوردار است اما در «پیکره» تدابیر خاصی جهت این موضوع اندیشیده شده و همگی واحدهای درون یک قطعه، برچسب‌گذاری اجزاء کلام شده‌اند. ازاین‌رو، کار قلاب‌گذاری در بیشتر موارد با مشکل جدی روبرو نخواهد شد. یکی از موارد اشکال برخی حروف اضافهٔ مرکب مانند «بدلیل، بخاطر» می‌باشند که به اشتباه به عنوان قطعهٔ چندواحدی در نظر گرفته شده و دو برچسب «حرف اضافه» و «اسم،عام،مفرد» گرفته‌اند. اما درمورد حالت دوم، موضوع متفاوت است و به این دلیل در مقالهٔ حاضر به این حالت پرداخته‌ایم.

۳-۲- واحدهای چندقطعه‌ای و انواع آن

اگر چند قطعه با هم تشکیل یک واحد زبانی را بدهند آن را واحدچندقطعه‌ای می‌نامیم [۸]. واحدهای چندقطعه‌ای در واقع واژه‌هایی هستند که به لحاظ دستوری باید دارای یک برچسب اجزای کلام باشند اما به دلیل آنکه این واژه‌ها از چند قطعه تشکیل شده‌اند و در نوشتار فارسی با نویسهٔ فاصله از هم جدا شده‌اند، هر کدام از این قطعه‌ها به تنهایی می‌تواند یک واژهٔ مجزا با برچسبی مختص به خود باشد. در پیکره اگر قطعه‌های یک واحد چندقطعه‌ای به صورت چسبیده به هم نوشته باشند و یا بنا به توصیه فرهنگستان زبان وادب فارسی، با نویسهٔ فاصلهٔ مجازی(نیم‌فاصله) به هم متصل شده باشند در آنصورت برچسب اجزای کلام واحدی گرفته‌اند مانند واژه‌های «بنابراین»، «دراین‌صورت»، «به هیچ وجه» که برچسب «حرف ربط» گرفته اند، «به‌سختی» و «به‌اشتباه» که برچسب «قید، کلی» گرفته اند. اما اگر قطعه‌ها با فاصله از هم نوشته شده باشند در آن صورت هر قطعه برچسب مجزایی گرفته است مانند «بنا بر این» که به عنوان سه واژه تلقی شده و سه برچسب «اسم،عام،مفرد»، «حرف‌اضافه» و «ضمیر،اشاره،مفرد» گرفته، «در این صورت» که سه برچسب «حرف‌اضافه»، «حرف‌تعریف،اشاره» و «اسم،عام،مفرد» دریافت نموده و «به سختی» و «به اشتباه» دو برچسب «حرف‌اضافه» و «اسم،عام،مفرد» گرفته اند.

در برخی موارد، قطعه‌ها به‌درستی به عنوان یک واحد در نظر گرفته شده‌اند و آن مواردیست که برخی از قطعه‌ها به عنوان تکواژهای مقیدی هستند که به تنهایی نمی‌توانند در متن ظاهر شوند و نقش نحوی مجزایی بگیرند مانند پسوند جمع، پیشوند استمراری و نیز تکواژهایی چون «نویس»، «دار»، «کار» و «فروش» و یا حالاتی که قطعه‌های کنار هم در اکثر موارد تشکیل یک واحد زبانی را می‌دهند مانند «دوچرخه سوار»، «بالا پوش»، «زیر پیراهنی».

در برچسب‌گذاری اجزای کلام پیکره، قطعه‌هایی که به عنوان تکواژهای مقید هستند، جداگانه برچسب‌گذاری نشده‌اند و طبعاً به درستی جزئی از قطعهٔ دارای برچسب مقولهٔ اصلی در نظر گرفته شده‌اند. ازاین‌رو، در مقالهٔ حاضر ما به بررسی این موارد که در پیکره به درستی تعیین وضعیت شده‌اند، نخواهیم پرداخت.

قبل از آنکه به دشواری مسالهٔ برچسب‌گذاری واحدهای چندقطعه‌ای در پیکره‌های زبانی بپردازیم، لازم است که حالتهای اصلی قرار گرفتن قطعه‌ها در کنار هم را بررسی کنیم. چهار حالت برای چند قطعهٔ کنارهم‌آمده می‌توان قائل شد.

حالت اول اینکه قطعه‌ها هریک دارای نقش صرفی و معنا هستند و همچنین مجموع قطعه‌ها تشکیل یک گروه نحوی را می‌دهند که معنای گروه، ترکیب معنای قطعه‌های درون آن است یا به اصطلاح یک «آرایش ترکیبی[۲۷]» است مانند «در کتاب دستورزبان» که یک گروه حرف‌اضافه‌ای است با یک معنای ترکیبی و «سیب سرخ خورشید» که یک گروه اسمی با معنایی ترکیبی می‌باشد.

حالت دوم اینکه ممکن است قطعه‌ها درمجموع یک گروه نحوی را تشکیل دهند که آن گروه دارای یک نقش صرفی و نیز یک معنا باشد که از جمع معنای قطعه‌ها حاصل نیامده باشد یا به اصطلاح یک «آرایش غیرترکیبی[۲۸]» باشد مانند افعال مرکب «دست شستن»، «داد زدن»، «کتک خوردن» و قیدهای «به ناگاه»، «به مراتب» و «به تدریج».

حالت سوم آنکه قطعه‌ها تشکیل یک گروه نحوی را بدهند که هم معنای ترکیبی داشته باشند و هم معنای غیرترکیبی. در حالت غیرترکیبی این

بسته‌ای را در فارسی تشکیل می‌دهند اختصاص دارد و در آن بخش در مورد مقوله‌های اصلی فعل و مصدر، حرف‌اضافه و ربط، قید، صفت و اسم، صحبت می‌کنیم. بخش ۵ بررسی واحدهای چندقطعه‌ای پویا را دربر می‌گیرد. بخش پایانی شامل نتیجه‌گیری است.

۲-معرفی «پیکره» و مجموعهٔ برچسب‌ها

پیکره دارای بیش از یکصد و ده میلیون کلمه شامل حدود ۳۶۰۰۰ پروندهٔ متنی است و هر پرونده شامل یک متن کامل یا نمونهٔ تصادفی منتخب از یک متن کامل می‌باشد که بر اساس معیارهای غیرزبانی و همچنین تعلق‌شان به گونه‌های فارسی معاصر انتخاب شده‌اند. هر متن رشته‌ای از جملات ساده یا مرکب است که با علائم سجاوندی از هم جدا شده‌اند. در انتخاب داده‌های پیکره دو نوع معیار نقش داشته‌اند: معیار زبانی و غیرزبانی. معیار زبانی عبارتست از پنج گونه زبان فارسی معاصر که بر حسب پارامترهای «معیار بودن» و «رسمی بودن» تعریف شده‌اند. علاوه بر آن، ده میلیون کلمه از متون پیکره به صورت تصادفی انتخاب و در سطح اجزاء کلام برچسب‌دهی شده‌اند.

در پیکره‌ها با توجه به ملاحظات خاص زبان مجموعه‌های مختلفی از برچسب‌ها استفاده شده‌اند[۸]. در تعریف برچسب‌های صرفی-نحوی پیکره چهار موضوع در نظر گرفته شده است: تعریف واژه، ساخت‌واژه، هم‌نگاره‌ها و اهداف.

در طراحی مجموعه برچسب‌های پیکره از راهنمای اروپایی ایگلز[۱۰] استفاده شده است. این راهنما برای نشانه‌گذاری دستوری[۱۱] متون زبان‌های اروپایی طراحی شد اما متخصصان پردازش زبان‌های طبیعی از آن برای برچسب‌دهی متون غیراروپایی نیز استفاده کردند. یکی از اهداف انتخاب این راهنما برای زبان فارسی، فراهم کردن زمینهٔ لازم برای ایجاد بانک درخت فارسی بوده است.

در راهنمای برچسب‌های ایگلز، سیزده مقولهٔ اجباری تعریف شده‌اند که درواقع معادل همان اجزای کلام در دستور سنتی می‌باشند. برای هر مقولهٔ اصلی، حداکثر سه مقولهٔ فرعی در نظر گرفته شده که هر مقوله فرعی شامل تعدادی ارزش یا مقدار است. مقوله‌های فرعی عبارتند از: ۱- مقوله‌های توصیه‌شده مانند نوع، جنسیت، شمار، زمان ۲- مقوله‌های اختیاری ذاتی که عمدتاً مشخصه‌های معنایی هستند مانند شمارش‌پذیربودن، جداپذیربودن، زمان، مکان و... ۳- مقوله‌های اختیاری خاص زبان مانند کسرهٔ اضافه، یاء نکره و پی‌چسب‌های ضمیری [۹].

در برچسب‌گذاری پیکرهٔ متون زبان فارسی معاصر از ۱۴ برچسب برای مقوله‌های اصلی، ۵۲ برچسب برای مقوله‌های توصیه‌شده، ۲۵ برچسب برای مقوله‌های اختیاری ذاتی و ۱۸ برچسب برای مقوله‌های اختیاری خاص زبان و در مجموع ۱۰۹ برچسب استفاده شده است. در نام‌گذاری برچسب‌ها از برچسب‌های کمک‌حافظه[۱۲] استفاده شده است. در ساختار نام کامل هر برچسب ازچپ به راست به ترتیب برچسب مقولهٔ اصلی، توصیه‌شده، ذاتی و خاص زبان آورده شده‌اند. [۱۰].

همانگونه که در بالا اشاره شد بر اساس راهنمای ایگلز، برچسب‌های صرفی-نحوی فارسی مشخص شده‌اند ولی به دلیل وجود شواهدی در زبان فارسی از برخی مشخصات ایگلز عدول شده است مثلاً ضمایر و حروف تعریف به عنوان دو مقولهٔ اجباری در نظر گرفته شده‌اند. همچنین حروف اضافهٔ پیشین و پسین نیز دو مقولهٔ جدا هستند در حالیکه در ایگلز یک مقولهٔ پوششی به نام حرف اضافه[۱۳] وجود دارد که دارای یک مقولهٔ توصیه‌شده «حرف اضافه پیشین» بوده و «حرف اضافه پسین» به عنوان مقولهٔ اختیاری خاص زبان در نظر گرفته می‌شود.

مقوله‌های اصلی تعریف‌شده در پیکرهٔ زبان فارسی معاصر به شرح زیر می‌باشند:

اسم (N)، فعل (V)، صفت(ADJ)، ضمیر(PRO)، توصیف‌گر(حرف تعریف) (DET)، قید(ADV)، حرف اضافه پیشین (PREP)، حرف اضافه پسین (POSTP)، حرف ربط (CONJ)، عدد(NUM)، حرف صوت(INT)، متفرقه(RES)، شاخص(رسته‌گر) (CL)، علایم سجاوندی(PUNC).

مقوله‌های اصلی فوق به عنوان واحدهای اساسی سازندهٔ گروه‌های نحوی می‌باشند.

همانگونه که گفته شد، بخشی از مقوله‌های فرعی، مقوله‌های توصیه‌شده هستند. به عنوان نمونه، برای اسم، مقوله‌های عام(COM) و خاص(PR) و نیز مفرد(SING) و جمع(PL)، در مورد صفت، ساده(SIM)، تفضیلی(COMP) و عالی(SUP)تعریف شده است. در مورد فعل نیز مقوله‌های توصیه‌شده‌ای معین گردیده است ازجمله **وجه:** التزامی(SUB)، امری(IMP)، **زمان:** حال(PRES)، گذشته(PA)، آینده(FUT). (برای اطلاعات و جزئیات بیشتر در مورد مقوله‌ها رجوع شود به [۱۰]).

در این بخش به معرفی اجمالی پیکره و مجموعه برچسب‌های اجزاء کلام تعریف‌شده در آن پرداختیم. در بخش بعد به مسائل خط فارسی و تعریف واحد چندقطعه‌ای خواهیم پرداخت.

۳-مسائل خط فارسی و واحدهای چندقطعه‌ای

در این بخش ابتدا به بررسی اجمالی مسائل خط فارسی که مرتبط با بحث واحدهای چندقطعه‌ای نیز می‌باشد پرداخته و سپس واحد چندقطعه‌ای را تعریف کرده و انواع ایستا و پویای آن را توضیح می‌دهیم.

۳-۱- خط فارسی و مسائل آن

الفبای عربی بعد از لاتین دومین رتبهٔ گستردگی نظام‌های نگارشی در جهان را داراست. زبان‌های غیرسامی چون فارسی، پشتو، اردو، سیندی، ویگور، کشمیری،کردی و آذربایجانی(در ایران) آن را به کار گرفته‌اند. برخی زبان‌ها مانند فارسی، متناسب با واج‌هایشان، تغییراتی در الفبا داده و یا حروفی را به آن اضافه نموده‌اند. الفبای زبان فارسی که پنج حرف اضافه‌تر «پ»،«ژ»، «گ»، «چ» و «ء» را نسبت عربی دارد، به نام الفبای فارسی-عربی[۱۴] نیز شناخته می‌شود. خط فارسی از راست به چپ به صورت سرهم نوشته می‌شود. منظور از سرهم نوشتن آن است که اکثر حروف یک واژه به هم می‌چسبند. الفبای فارسی یک «*ابجد ناخالص*[۱۵]» است که در آن واکه‌های بلند (/ ا/ و /i/ و /u/) نوشته می‌شوند ولی واکه‌های کوتاه (/æ/، /e/ و /o/)معمولاً در خط نوشته نمی‌شوند. این باعث می‌شود که خواندن خط فارسی دشوار گردد زیرا آنچه نوشته می‌شود با آن آوایی که از آن بدست می‌آید متفاوت است. بنا به این دلیل از ۱۱ «نشانهٔ زیر و زبری[۱۶]» به صورت اختیاری و در برخی موارد اجباری که در بالا و پایین حروف نوشته می‌شوند جهت حل مسالهٔ بازسازی تلفظ واژه‌ها و ابهام‌زدایی استفاده می‌گردد. ۲۶ حرف فارسی با توجه به جایگاه‌شان در رشتهٔ حروف، دارای چهار صورت مجزا[۱۷]، آغازی[۱۸]، میانی[۱۹] و پایانی[۲۰] هستند[۱۱]. این حروف که حروف اتصالی نام دارند عبارتند از : ء، ب، پ، ت، ث، ج، چ، ح، خ، س، ش، ص، ض، ط، ظ، ع،

۱-مقدمه

در زبان‌های دارای نوشتار مبتنی بر خط عربی از جمله عربی، اردو، پشتو، کردی، آذری و فارسی، به‌دلیل صورت‌های نوشتاری مختلف اکثرِ حروف از جمله صورت‌های مجزا، آغازی، میانی و پایانی، واحدهای چندقطعه‌ای معمولاً به صورت سرهم یا جداازهم نوشته می‌شوند. در حالت جداازهم، یک واحد زبانی، معمولاً به عنوان چند واژه در نظر گرفته شده و طبعاً هر واژه برچسب‌های صرفی مختلفی نسبت به حالت سرهم می‌گیرد. این عدم یکپارچگی برچسب‌ها در سطح اجزای کلام و برچسب‌گذاری نادرست صرفی باعث بروز مشکلاتی جدی در تولید بانک درخت گروه‌های نحوی[۱] متون فارسی (از این به بعد به اختصار «بانک درخت فارسی» نامیده می‌شود) و نیز تحلیل‌کننده‌های صرفی و تجزیه‌گرهای نحوی می‌گردد. شناسایی و تجزیه وتحلیل پیکره‌بنیاد واحدهای چندقطعه‌ای در متون فارسی معاصر، به ایجاد یکی از دو واحد فرعی به نام «واحد تقطیع متون[۲]» درون نظام پردازش‌گر تولیدکنندهٔ بانک درخت فارسی می‌انجامد. در مقالهٔ حاضر، با بهره‌گیری از اطلاعات متون دارای برچسب اجزاء کلام[۳] در پیکرهٔ متنی زبان فارسی معاصر، به تجزیه و تحلیل واحدهای چندقطعه‌ای ثابت و پویا[۴]ی زبان فارسی در سطح اکثر مقوله‌های اصلی از جمله فعل، حروف ربط و اضافه، قید، صفت و اسم پرداخته و نشان میدهیم که هریک از این واحدهای چندقطعه‌ای چگونه در تولید بانک درخت فارسی تاثیرگذار هستند.

طراحی و پیاده‌سازی هر نظامی، مستلزم تعیین درون‌دادهای آن نظام، حوزه‌های پردازشگر درون آن نظام و نیز برون‌دادها می‌باشد. متون فارسی به همراه اطلاعات برچسب‌های اجزای کلام آنها به عنوان درون‌دادهای حوزهٔ تقطیع‌کننده و نتیجتاً نظام ایجادکنندهٔ بانک درخت فارسی می‌باشند. به بیان دیگر، متون برچسب‌گذاری‌شده در سطح واژه، به عنوان ورودی وارد یک نظام پردازشگر می‌شوند. درون آن نظام، با استفاده از واحدهای فرعی تقطیع‌کنندهٔ متون و قلاب‌گذار نحوی ، عمل گروه‌بندی بر روی متون فارسی صورت می‌گیرد. بانک درخت فارسی، به عنوان برون‌داد این واحدها و در نتیجه این نظام پردازشگر، حاصل می‌آید.

حال به‌طوراجمال به مطالعات پیشین درخصوص بانک درخت گروه‌های نحوی می‌پردازیم. در سال ۱۹۶۱ میلادی، پیکرهٔ آمریکایی براون با حجم بیش از یک میلیون کلمه در دانشگاه براون با هدف مطالعهٔ آماری زبان انگلیسی آمریکایی به صورت الکترونیکی تولید شد و در اختیار محققان قرار گرفت[۱]. پیکرهٔ زبانی براون که پیشگام در این موضوع می‌باشد، کاربرد گسترده‌ای در مطالعات زبان‌شناسی رایانه‌ای داشته است.

پس از آن، بانک درخت پن انگلیسی با بیش از ۴/۵ میلیون واژه بر اساس پیکرهٔ براون بنا شده و اساس کارهای بعدی قرار گرفته است. هدف پروژهٔ بانک درخت پن، برچسب‌گذاری در سطح واژه و پس از آن در سطح گروه نحوی بوده است. گفتنی است که برچسب‌گذاری صرفی و نحوی به‌صورت نیمه‌خودکار انجام پذیرفته است. یعنی ابتدا متون با یک برنامهٔ رایانه‌ای به نام تجزیه‌گر فیدیچ برچسب‌گذاری شده‌اند و سپس خطاها به صورت دستی از طریق یک واسط گرافیکی ویرایش‌گر مورد اصلاح قرار گرفته‌اند.

بانک درخت پن با مجموع ۴۸ برچسب نسبت به پیکرهٔ براون با ۸۷ برچسب تعداد برچسب‌های کمتری دارد. انجام چنین کاری بنا به دلایل بازیافتنی‌بودن[۵] و نیز یکپارچگی[۶] برچسب‌ها بوده است [۲].

همچنین در بانک درخت پن بر خلاف پیکرهٔ براون به نقش دستوری کلمات نیزتوجه می‌شود. به این معنی که یک واژه ممکن است برچسب اجزای کلام متفاوتی بگیرد مانند کلمهٔ both که می‌تواند برچسب‌های پیش‌توصیف‌گر، قید، اسم عام جمع و حرف ربط همپایه‌ساز را بسته به بافت نحوی‌اش دریافت نماید. مجموعهٔ برچسب‌های نحوی بانک درخت پن شامل ۱۴ برچسب می‌باشد که از آن میان می‌توان به گروه اسمی (**NP**)، گروه صفتی(**ADJP**)، گروه حرف‌اضافه‌ای(**PP**)، گروه فعلی (**VP**)، بند خبری(**S**) و بند متممی (**SBAR**) اشاره نمود. همچنین ۴ برچسب جهت عناصر تهی شامل فاعل گروه بدون زمان(*)، متمم‌نمای تهی (**0**) و ردّ (**T**) در نظر گرفته شده است.(برای جزئیات بیشتر به [۳] مراجعه نمایید).

درقلاب‌گذاری گروه‌های نحوی بانک درخت پن، علاوه بر مشخص شدن مرز گروه‌های نحوی، هر گروه ممکن است برچسب‌های دیگری نیز بگیرد که مربوط به نقش‌های نحوی، معنایی، مبتدایی و غیره باشد.

دسته‌ای از این برچسبهای نقشی مربوط به عبارات قیدی می‌باشند که معمولاً ادات گروه فعلی هستند مانند عبارات قیدی جهت (**DIR-**)، محل و موقعیت یک رویداد (**LOC-**)، عبارات قیدی نشان‌دهندهٔ حالت (**MNR-**) و زمان(**TMP-**)، فاعل منطقی در جملات مجهول(**LGS-**)، گزاره(**PRD-**) و منادا (**VOC-**). در بخش ۳ به برخی از این نوع برچسب‌های نقشی می‌پردازیم.

بر اساس بانک درخت پن انگلیسی، بانک درخت‌های دیگری نیز به وجود آمده‌اند که می‌توانیم به بانک درخت پن عربی و بانک درخت عبری اشاره کنیم. در این بانک درخت‌ها، ابتدا متون در سطح تکواژها تحلیل صرفی و برچسب‌گذاری شده و سپس مورد برچسب‌گذاری نحوی به صورت نیمه‌خودکار قرار گرفته‌اند. بااینکه چارچوب اصلی کارشان همان بانک درخت پن انگلیسی بوده ولی متناسب با شرایط خاص زبان عربی و عبری، تغییراتی در نحوهٔ برچسب‌گذاری صرفی و گروه‌بندی ایجاد نموده‌اند. مثلاً در عربی فاعل درون گروه فعلی در نظر گرفته شده است و یا در بانک درخت عبری، بدلیل غنای بیشتر صرف آن نسبت به انگلیسی، مشخصه‌های تطابق برای جنسیت، شمار، شخص و زمان در بسیاری از برچسب‌های اجزای کلام آورده شده است. [۴]و [۵].

در ایران دو مرکز تحقیقاتی وجود دارد که پیکره‌های گفتاری و نوشتاری تولید کرده‌اند که بیشتر در داخل ایران استفاده می شود: پژوهشگاه علوم انسانی و مطالعات فرهنگی و پژوهشکدهٔ پردازش هوشمند علائم.

پژوهشگاه علوم انسانی و مطالعات فرهنگی یک دادگان برخط به نام «دادگان زبان فارسی» تولید نموده است که در آن میلیون‌ها واژه شامل متون زبان و ادبیات فارسی معاصر از انواع مختلف جمع‌آوری شده و سپس حجم کوچکی از آن به صورت نمونه با استفاده از ۴۵ برچسب صرفی نشانه‌گذاری شده استفاده شده است. متون شامل هر دو سبک رسمی و غیررسمی است[۶]. در طراحی این برچسب‌ها از معیارهای پیشنهادشده در [۷] استفاده شده است.

آخرین پیکره‌ای که پژوهشکدهٔ پردازش هوشمند علائم تولید نموده است، «پیکرهٔ متنی زبان فارسی معاصر»(از این به بعد به اختصار «پیکره»[۷] نامیده می‌شود) می‌باشد.

مقالهٔ حاضر بدین صورت بخش‌بندی شده است که در بخش ۲ به معرفی اجمالی «پیکره» و مروری بر انواع برچسب‌های اصلی و فرعی تعریف‌شده در آن خواهیم پرداخت. ازآنجاییکه در تجزیه و تحلیل واحدهای چندقطعه‌ای بیشتر با مقوله‌های اصلی و نیز بخشی از مقوله‌های فرعی سروکار داریم، تأکیدمان بیشتر برروی این نوع برچسب‌ها خواهد بود. در بخش ۳ مروری گذرا بر مسائل خط فارسی که مبتنی بر خط عربی است نموده و سپس به تعریف و توضیح واحد چندقطعه‌ای[۸] و انواع ایستا[۹] و پویای آن می‌پردازیم. بخش ۴ به واحدهای چندقطعه‌ای ایستا که طبقهٔ

مجله بین المللی
ارتباطات و فناوری اطلاعات

ITRC

جلد یکم - شماره ۳ - تیر ماه ۱۳۸۸ (ش ص . ۲۶-۱۵)

تجزیه و تحلیل پیکره‌بنیاد واحدهای چندقطعه‌ای در متون فارسی

مسعود شریفی آتشگاه

گروه زبان‌شناسی
دانشکدهٔ ادبیات و علوم انسانی
دانشگاه تهران
m_sharifi@nigc-tpgc.ir

محمود بی‌جن‌خان

گروه زبان‌شناسی
دانشکدهٔ ادبیات و علوم انسانی
دانشگاه تهران
mbjkhan@ut.ac.ir

تاریخ دریافت: ۱۳۸۸/۱/۲۹- تاریخ پذیرش: ۱۳۸۸/۴/۲۶

چکیده - **برچسب‌گذاری صرفی و نحوی واحدهای چندقطعه‌ای به دلیل سرهم بودن نوشتار فارسی و لذا وجود تنوع نوشتاری با مشکلات متعددی همراه است. در مقالهٔ حاضر با بررسی خط فارسی و مسائل مربوط به آن، آرایش‌های ترکیبی ، غیرترکیبی و نیز آرایش جدید نیمه‌ترکیبی که صورت‌های همنشینی قطعه‌ها می‌باشند توضیح داده می‌شوند. سپس به منظور تبیین این آرایش‌ها، واحدهای چندقطعه‌ای ایستا و پویا برای ساختارهای غیرزایا و زایای هریک از مقوله‌های اصلی از جمله فعل و مصدر، حرف‌اضافه و ربط، قید، صفت و اسم تشریح می‌شوند. ایجاد الگوهای واحدهای چندقطعه‌ای برای این مقوله‌ها از نتایج مهم این تحقیق می‌باشد. همچنین نتایج حاصل می‌توانند به عنوان درون‌دادهای حوزهٔ تقطیع‌کننده در نظام تولیدکنندهٔ بانک درخت گروه‌های نحوی مورد استفاده قرار گیرند. کاربرد دیگر این تحقیق در طراحی تحلیل‌کننده‌های صرفی و نیز تجزیه‌گرهای نحوی می‌باشد.**

واژه‌های کلیدی - برچسب اجزاء کلام، واحد چندقطعه‌ای، ایستا، پویا، بانک درخت گروه‌های نحوی، تقطیع متون، آرایش ترکیبی، غیرترکیبی، نیمه‌ترکیبی، الگو

***Abstract*- Because of the joining behavior of Persian script and its orthographic variation, the morphological and syntactic annotations of multi-token units meet various issues. By the analysis of Perso-Arabic script and its problems, the various collocation types of the tokens including the compositional, non-compositional and the new semi-compositional constructions are described in the present paper. Then, to illustrate these constructions, the static and dynamic multi-token units will be presented for the generative and non-generative structures of the main categories including the verbs, infinitives, prepositions, conjunctions, adverbs, adjectives and nouns. Defining the multi-token unit templates for these categories is one of the important results of this research. The findings can be input to the segmentation module of the Persian Treebank generator system. The other usage of the present research is in the design and implementation of the morphological analyzers and syntactical parsers.**