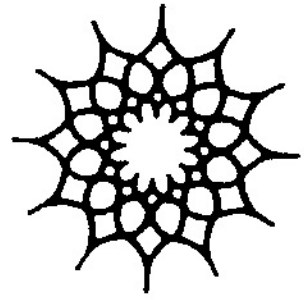


پژوهشگاه علوم انسانی و مطالعات فرهنگی

پژوهشکدهٔ زبانشناسی

پایان‌نامه

برای دریافت درجه کارشناسی ارشد زبانشناسی همگانی

موضوع

برچسب‌دهی دستوری پیکرهٔ زبان فارسی

پژوهشی در زبانشناسی رایانه‌ای

استاد راهنما: دکتر سید مصطفی عاصی

استاد مشاور: دکتر شاپور اعتماد

پژوهش از: محمد حاجی عبدالحسینی

اسفند ۱۳۷۵

فهرست مندرجات

پیشگفتار

گسترش سریع و فزایندهٔ رایانه و بسط بی‌حد و حصر امکانات این ابزار قدرتمند، موجب گردیده است که امروزه تقریباً تمام شاخه‌های علوم و هنر از آن بهره‌مند گردند. زبانشناسی نیز از این دسته مستثنی نیست. اتفاقاً زبان جزء اولین حوزه‌هایی بود که کاربرد رایانه در آن مدّ نظر قرار گرفت. منظور دورهٔ تب ساختن ماشین‌های ترجمه در دهه‌های ۱۹۵۰ و ۱۹۶۰ است. صحیح است که پیچیدگی زبان و سادگی رایانه‌های ابتدایی و نیز ضعف نظریه‌های زبانی آن دوران موجب رکود فعالیت‌های زبانشناسی رایانه‌ای گردید، اما از دههٔ ۱۹۷۰ این گونه فعالیت‌ها و مطالعات با بینشی نو از سرگرفته شد. در این زمان زبانشناسی رایانه‌ای به حدی گسترش یافته است که صحبت از مترجم‌های شفاهی ماشینی، سیستم‌های یادگیرنده، تولید متن به وسیله رایانه و بسیاری فعالیتهای دیگر به میان آمده است.

اما زبانشناسی رایانه‌ای در کشور ما هنوز بسیار ناشناخته باقی مانده و چشم‌اندازهای وسیع نظری و کاربردی آن از دید عدهٔ کثیری به دور است. نگارنده به دلیل علاقهٔ فراوان به شاخهٔ زبانشناسی رایانه‌ای برآن شد تا پایان‌نامهٔ کارشناسی ارشد خود را در این زمینه انتخاب نماید. این کار هم می‌توانست به آشنایی بیشتر نگارنده به این رشته و هم به شناساندن آن به جامعهٔ زبانشناسی ایران کمکی هر چند ناچیز کرده باشد. استاد راهنمای محترم نگارنده، جناب آقای دکتر عاصی، ضمن ارائه رهنمودهای گوناگون، پایگاه داده‌های زبان فارسی و اهداف آن را معرفی نموده و پیشنهاد کردند که چه بهتر موضوع یک پایان‌نامه تنها پایان کار نبوده بلکه کاربردهای عملی در یک پروژهٔ وسیع‌تر داشته باشد و چه قدر این امر صحیح است. بدین ترتیب، نگارنده موضوع برچسب‌دهی دستوری پیکرهٔ زبان فارسی راکه یکی از مراحل ایجاد و گسترش دامنهٔ کاربرد یک پیکرهٔ زبانی است انتخاب نمود.

هدف از یک طرح برچسب‌دهی، مشخص کردن مقولهٔ دستوری هر یک از اقلام واژگانی موجود در یک متن است. درگذشته چنین کاری به طور دستی انجام می‌گرفت، اما اکنون امکان انجام خودکار آن با استفاده از رایانه پیدا شده است. البته طبق معمول

پیچیدگی‌های زبان موجب می‌گردد که کار آن طورکه به نظر می‌رسد ساده نباشد. به همین دلیل است که، روش‌های متفاوتی برای انجام این کار در کشورهای دیگر ابداع شده است.

از آنجاکه این کار برای اولین بار روی زبان فارسی انجام می‌پذیرد و برای آن که بتوان سیستم را در مدت زمان تعیین شده برای انجام پایان‌نامه فعال نمود، تصمیم گرفتم طرحی ساده و نیمه خودکار را برای این کار انتخاب نمایم. عمدهٔ کار این پایان‌نامه صرف نگارش برنامهٔ اصلی پروژه گردید. روشی که در برنامه اتخاذ شده است، عمدتاً متکی است بر روش معرفی شده در شوتس (۱۹۹۵) ولی برای این که با محدودیت‌های انجام این پروژه سازگار شود، تعدیل‌هایی در روش مزبور اعمال شده است که طبیعتاً از دقت بخش خودکار سیستم می‌کاهد. بخش دوم کار مربوط بود به تدوین یک مجموعهٔ برچسب که با استفاده از آن بتوان متن را برچسب‌دهی نمود. مجموعهٔ برچسب باید به‌گونه‌ای طراحی می‌شد که با روش به‌کار رفته در پروژه سازگار باشد و نیز از معیارهای عمومی برای ایجاد چنین مجموعه‌ای تخطی نکند. بخش سومِ کار، به آزمایش پروژه و بررسی نتایج اختصاص یافت.

پایان‌نامهٔ حاضر شامل پنج فصل است. فصل اول به معرفی رشتهٔ زبانشناسی پیکره بنیاد می‌پردازد که برچسب‌دهی دستوری شاخه‌ای از آن است. در این فصل گفته می‌شود که در زبانشناسی پیکره بنیاد یکی از اعمال اصلی که روی پیکره انجام می‌شود، نشانه‌گذاری و آماده کردن آن برای کاربردهای گوناگونی است که می‌تواند داشته باشد. در فصل دوم انواع نشانه‌گذاری معرفی می‌گردد. فصل سوم اختصاص دارد به برچسب‌دهی دستوری که نوع خاصی از نشانه‌گذاری پیکره و موضوع اصلی این پروژه است. در فصل مذکور روش اصلی در این کار معرفی می‌گردد و روش اتخاذ شده شرح داده می‌شود. فصل چهارم که بخش عمده‌ای از پایان‌نامه را به خود اختصاص داده است، شامل معرفی پروژهٔ برچسب‌دهی دستوری فارسی است. در بخش اول آن معیارهای ایجاد یک مجموعهٔ برچسب و سپس مجموعهٔ برچسب طراحی شده برای پروژه ارائه می‌گردند و در بخش دوم آن برنامه‌های نوشته شده و برنامه‌های به کار رفته در پروژه تک تک معرفی و شرح داده می‌شوند. فصل پنجم، شرح آزمایش‌های انجام شده روی بسته نرم‌افزار و بررسی نتایج به دست آمده را شامل

شده و در انتهای فصل مزبور پیشنهادهایی برای پروژه‌های بعدی مطرح می‌شود.

در اینجا جا دارد از کسانی که نگارنده را در به انجام رسانیدن این پروژه یاری کرده‌اند قدردانی شود. ابتدا از جناب آقای دکتر سید مصطفی عاصی تشکر می‌کنم که با گشاده‌رویی و با شکیبایی فراوان مزاحمت‌های مکرر مرا تحمل کرده و از ارائه هیچ‌گونه کمکی دریغ ننموده‌اند. زبانشناسی تنها گوشه‌ای از مطالبی است که بنده از ایشان فراگرفته‌ام. دیگر از جناب آقای دکتر شاپور اعتماد متشکرم که علیرغم گرفتاری‌های فراوان و ضیق وقت مشاورهٔ کار را برعهده گرفتند. از برادرم آقای علی حاجی عبدالحسینی بی‌نهایت سپاسگزارم که دستگاه رایانهٔ خود را در اختیارم قرار داده و مزاحمت‌های ممتد مرا تا پاسی از شب گذشته متحمل می‌شدند. از پرسنل مرکز کامپیوتر پژوهشگاه علوم انسانی و مطالعات فرهنگی به دلیل همکاریهایشان ضمن جستجو برای منابع و نگارش، غلط گیری و چاپ برنامه‌ها متشکرم. در پایان از سرکار خانم محجوب که زحمت تایپ پایان‌نامه را متقبل شدند و سرکار خانم صفوی که امکان چاپ رایانه‌ای آن را برای نگارنده فراهم ساختند بی‌نهایت سپاسگزارم.

محمد حاجی عبدالحسینی

۱۳۷۵/۱۱/۲۴

فصل ۱ زبانشناسی پیکره‌بنیاد

۱ـ۱ درآمد

زبانشناسان همواره بر بنیاد نهادن بررسی‌های زبانی و استدلال‌ها و استنتاج‌های مبتنی بر داده‌های واقعی زبان تأکید داشته و دارند، و همواره بهترین راه برای استفاده از داده‌های مستند را ایجاد و استفاده از پیکره‌های زبانی می‌دانسته‌اند. از طرفی گسترش روزافزون رایانه و علوم مختلف میان رشته‌ای، این امکان را برای زبانشناس فراهم می‌آورد که بتواند پیکره‌های بسیار عظیم زبانی را بوجود آورده، آنها را به گونه‌های مختلف طبقه‌بندی و نشانه‌گذاری نموده و به طرق بسیارگوناگون از اطلاعات به دست آمده بهره‌مند گردد.

گنجایش سیستم‌های ذخیره‌سازی اطلاعات و سرعت پردازش رایانه‌های امروزی به حدّی بالا رفته است که به پژوهشگر امکان می‌دهد تا کاری را که گاهی ممکن بود ماه‌ها به طول انجامد و مخارج زیادی در بر داشته باشد، ظرف مدت چند دقیقه و بادقت بالا به انجام رساند. به همین دلیل، در طی دهه‌های ۱۹۸۰ و ۱۹۹۰ شاخه‌ای جدید از زبانشناسی شروع به رشد کرد که امروزه آن را با نام «زبانشناسی پیکره بنیاد»[1] می‌شناسیم. «گرایش فعلی در زبانشناسی پیکره بنیاد به سوی ایجاد پیکره‌های متنی هر چه بزرگتر است» (پیچی[2]، ۱۹۹۴: ۵۰۱). هُلمس - هیگین[3] و دیگران (۱۹۹۴: ۳۹۰) اظهار می‌دارند که «استفاده از پیکره‌های متنی، بویژه استفاده از پیکره‌های متنی رایانه‌ای، کاربردهای مفید ویژه‌ای برای مطالعهٔ زبان‌ها و شاید در حدی پایین‌تر برای آموزش و یادگیری زبان داشته است.» از میان کاربردهای متعدد پیکره‌های زبانی می‌توان به اشاره به موارد زیر بسنده کرد: پیکره‌ها ممکن است؛

۱ـ منبعی باشند برای پژوهش‌های زبانشناختی و آموزش زبان،

۲ـ منبعی باشند برای فرهنگ‌نگاری،

۳ـ درسیستم‌های درک گفتار طبیعی به کار روند،

۴ـ در سیستم‌های ترجمه براساس اطلاعات آماری به کار روند، و یا

---

1. Corpus Linguistics 2. Picchi 3. Holmes-Higgin

۵ـ منبعی باشند برای ارزیابی و استنتاج خودکار یا نیمه خودکار دستورهای زبان (گارساید[۴]، ۳۹:۱۹۹۳).

واضح است که با وجود پیکره‌های زبانی عظیم و در حال رشد و با وجود کاربردهای بسیار گوناگون برای این پیکره‌ها، نمی‌توان آنها را به شکل خام و مانند مجموعه‌ای ساده از متون مختلف نگهداری نمود، بلکه این متون می‌باید دسته‌بندی و به روش‌های مختلف نشانه‌گذاری گردند تا بتوانند پاسخگوی نیازهای کاربران باشند. در واقع ایجاد یک پیکرهٔ متنی رایانه‌ای یعنی تبدیل متون منتشر شده به یک بانک اطلاعاتی. هُلمس - هیگین و دیگران (۳۹۰:۱۹۹۴) مراحل زیر را برای این تبدیل می‌شناسند:

۱ـ *رمزگذاری*[۵]

۲ـ *توصیف*[۶] متن

۳ـ و در صورت امکان *بازنمایی*[۷] متن

رمزگذاری الکترونیک متن، در اصل عبارت است از علامت‌گذاری قسمت‌های متفاوت متن از جمله اطلاعات مربوط به پیکربندی آن و غیره، به طوری که بتوان اطلاعات مربوط به پیکربندی را از خود متن جدا نمود. در بخش ۱ـ۲ـ۱ در این مورد بیشتر توضیح داده خواهد شد. توصیف متن به این معنی است که محتوای متن‌های درون یک پیکره برای کاربران توضیح داده می‌شود. این کار خود نیازمند طبقه‌بندی متون موجود و ایجاد امکان دسترسی به متون براساس این طبقه‌بندی (ها)برای کاربران از طریق یک *برنامهٔ مدیریت پیکره*[۸] است.

مرحلهٔ سوم، یعنی بازنمایی متن، از دو مرحلهٔ پیشین بسیار پیچیده‌تر است و طی آن باید قراردادهای نحوی و معنایی خاصی را مشخص کرد که مطابق آنها بتوان متن را روی یک سیستم رایانه‌ای بازنمایی کرد. زمانی که یک متن در چنین سیستمی *بازنموده* شد، امکان این وجود خواهد داشت که یک نرم‌افزار دیگر بتواند اطلاعات تازه‌ای را از متن استنتاج کند؛ در واقع به تعبیری می‌توان گفت که رایانه خواهد توانست با استفاده

---

4. Garside
5. coding
6. description
7. representation
8. corpus management program

... قراردادهای نحوی و معنایی متن را درک کند (هُلمس - هیگین و دیگران، ۱۹۹۴: ۳۹۰-۳۷۰). ... برچسب‌گذاری و بازنمایی متن باید آن را طبق قراردادهای خاصی نشانه‌گذاری[9] کرد.

۲-۲ نشانه‌گذاری پیکره

جفری لیچ[10]، استاد زبان‌شناسی و زبان انگلیسی نوین در دانشگاه لنکستر[11] بریتانیا، در مقاله ۱۹۹۳ خود (ص ۲۷۵) نشانه‌گذاری پیکره را این چنین تعریف می‌کند:

«نشانه‌گذاری پیکره عبارت است از عمل افزودن اطلاعات تفسیری (بویژه اطلاعات زبانی) به یک پیکرهٔ گفتاری و /یا نوشتاری زبان، با استفاده از نوعی برچسب‌گذاری به شکل خودکار[12] الکترونیک ماده زبانی.»

از آنجا که «اطلاعات تفسیری» می‌توانند انواع مختلفی داشته باشند، پس انواع مختلف نشانه‌گذاری نیز وجود دارد (لیچ، ۱۹۹۳: ۲۷۸-۲۷۵):

۱- نگارشی[13]

۲- واژی-نحوی

۳- ساخت‌جزیی

۴- دستوری (برچسب‌های دستوری)[14]

۵- نحوی (تجزیه)[15]

۶- معنایی

۷- گفتارکرد (ساخت کلامی)

۲-۲-۱ نشانه‌گذاری نگارشی

نوع اول نشانه‌گذاری که لیچ معرفی می‌کند یعنی *نشانه‌گذاری نگارشی* در واقع همان برچسب‌گذاری متنی است که هُلمس - هیگین و دیگران به آن اشاره کرده‌اند (ر.ک. ۲-۱). این نشانه‌گذاری نگارشی ابتدا نیاز به یک *زبان علامت‌گذاری*[16] وجود دارد. بر همین اساس چنین زبانی جزء یک متن بر اساس نقش خود علامت‌گذاری می‌شوند. اجزاء متن می‌توانند هر بخشی از متن باشند مانند یک پاراگراف، چکیده، یادداشت و یا

---

9. annotation
10. Geoffrey Leech
11. Lancaster
12. representation
13. orthographic
14. grammatical tagging
15. parsing
16. markup language

شکل ۱-۱ نمونه‌ای از علامت‌گذاری SGML (اسمیت، ۱۹۸۷: ۱۷۲)

```
<sd
<mb
<h1 The Standard Generalized Markup Language (SGML) for
Humanities Publishing
<au Joan M Smith
<ad
<1>National Computing Centre
</ad>
<ab
<p>A new methodology, at the core of which is generic coding,
has been developed within the International Organization for
Standardization (ISO).
This is known as the &SGML;.
Using SGML, the elements of a document are marked up as to
their role, be it a paragraph, an abstract, a note, or
whatever  the style of presentation is a separate issue and
is not addressed by SGML.
These elements can form part of a data base, which can be
updated at will.
So there is the notion of data base publishing.
<p>The &SGML  is presented as a tool for full&en;text data
base publishing, where the options for output are open, an
example being given of a marked up document.
Its value for all aspects of humanities publishing is
addressed:  whether for scholarly papers intended for a
journal, books, specialist publications, dictionaries, or
bibliographies; indeed, whatever is input to an electronic
medium with the intention of its being imaged subsequently in
some form- whether alone or in combination with other text.
SGML represents an advance in publishing methodology, taking
advantage of developing technology.
It can be exploited as such in an academic environment to
give an added dimension to research publications.
```

هر عنصر دیگر (اسمیت[17]، ۱۹۹۰: ۱۰۱). هنگامی که متنی با استفاده از یک زبان علامت‌گذاری، نشانه‌گذاری نگارشی گردید، آنگاه می‌توان آن متن را بسادگی به رایانه‌های دیگر منتقل نمود.

از بین زبان‌های علامت‌گذاری می‌توان به SGML[18] که توسط سازمان استاندارد جهانی (ISO) ساخته شده و TEI[19] شاخه‌ای نمونه از آن است بین‌المللی که تحت حمایت مالی انجمن رایانه و علوم انسانی[20]، انجمن زبانشناسی رایانه‌ای[21] و انجمن زبانشناسی و ادبیات رایانه‌ای[22] فعالیت دارد.

شکل ۱-۱ نمونه‌ای از یک متن انگلیسی که با SGML نشانه‌گذاری شده نشان می‌دهد. همان‌طور که ملاحظه می‌شود نشانه‌ها یا علامت‌های < > قرار داده می‌شوند تا رایانه بسادگی بتواند آنها را از خود متن تشخیص دهد. به عنوان مثال در نمونهٔ مزبور <sd> به معنی شروع متن (start document) است و <mb> یعنی بدنهٔ اصلی متن (main body).

بدین ترتیب رایانه می‌تواند با هر یک از اجزاء متن به شکل متفاوتی برخورد نماید، مثلاً" بخشی را با قلم خاصی بنویسد و یا فقط قسمت خاصی از متن را نمایش داده یا به رایانهٔ دیگری منتقل نماید.

۱-۲-۲ نشانه‌گذاری آوایی /واجی

این نوع نشانه‌گذاری در علوم زبانی و زبانشناسی گستردگی گسترده‌ای دارد.

۱-۲-۳ نشانه‌گذاری زیر زنجیری

لیچ (۱۹۹۳: ۲۷۶) اظهار می‌دارد که نشانه‌گذاری شده از لحاظ ویژگی‌های زیر زنجیری وجود دارد: ۱- پیکرهٔ متنی لانکستر که شامل یکصدهزار واژهٔ انگلیسی

---

17. Joan M.Smith
18. Standard Generalized Markup Language
19. Text Encoding Initiative
20. The Association for Computers and Humanities
21. The Association for Computational Linguistics
22. The Association for Literary and Linguistic Computing
23. The London-Lund Corpus

بریتانیا می‌شود، ۲- پیکرهٔ انگلیسی گفتاری لنکستر/آی بی ام[۲۴]؛ این پیکره تنها شامل پنجاه هزار واژه است اما در سطوح مختلف زبانی نشانه‌گذاری شده است.

این گونه نشانه‌گذاری دستی انجام می‌شود و مهارت نشانه‌گذار در تشخیص ویژگی‌های زبرزنجیری درست، از اهمیت خاصی برخوردار است. شکل ۱-۲ نمونه‌ای از یک مکالمهٔ نشانه‌گذاری شده را به همراه معانی نشانه‌ها نمایش می‌دهد.

۱-۴-۴ برچسب دهی دستوری

این نوع نشانه‌گذاری که موضوع همین پایان‌نامه است در فصل‌های بعدی به تفصیل توضیح داده خواهد شد.

۱-۴-۵ تقطیع نحوی

غلب به برچسب‌های دستوری به عنوان مرحلهٔ اول یک نشانه‌گذاری نحوی جامع‌تر نگریسته می‌شود. این نشانه‌گذاری عبارت است از تهیهٔ یک نمایش درختی[۲۵] یا قلاب‌های نشاندار[۲۶] برای هر یک از جمله‌های درون پیکره. پیکره‌های تقطیع شدهٔ تهیه از بانک درختی[۲۷] می‌نامند. بانک‌های درختی گوناگون وکوچکی تهیه شده‌اند و به منظور افزایش سرعت و صحت نشانه‌گذاری گرایش به این سو پیدا شده است که از صورت‌های ساده‌تری استفاده شود. شکل‌های ۱-۳ و ۱-۴ دو نمونه از تقطیع جملات انگلیسی را نشان می‌دهند.

۱-۴-۶ نشانه‌گذاری معنایی

پس از نشانه‌گذاری نحوی (یعنی برچسب دهی دستوری و تقطیع جملات) روشن است که قدم بعدی نشانه‌گذاری معنایی است. برچسب دهی معنایی کلمات مثلاً می‌تواند با این هدف انجام شود که معنای واژگانی کلمات در متن شناسایی شوند؛ این روشی است که به آن تعیین معنی[۲۸] می‌گویند. شکل ۱-۵ نمونه‌ای است از برچسب‌دهی معنایی کلمات که از یک پروژهٔ تحلیل خودکار معنا در دانشگاه لنکستر گرفته شده است. در این مثال، متن از بالا به پایین خوانده می‌شود: برچسب‌های دستوری در سمت چپ قرار دارند و برچسب‌های معنایی در سمت راست.

---

24. The Lancaster/IBM Spoken English Corpus
25. phrase marker
26. labelled brackets
27. treebank
28. sense resolution

شکل ۱-۲ نشانه‌گذاری زبر زنجیری، پیکرهٔ لندن - لاند

توجه: البته این نسخهٔ ASCII می‌تواند برای کاربر نهایی (ترمز) تسهیل (لیچ، ۱۹۹۳: ۲۷۶)

well ^very nice of you to ((come and)) _spare the
:time and#
^come and !t\alk# -
^tell me a'bout the - !pr\oblems#
and ^inci\dentally# .
^ :@:| ^do ^do t\ell me#
^anything you 'want about the :college in "!g\eneral#
' ^mean it !doesnt "!h\ave to be (con^f\ined#)#
to the ^problems of !\English#
:@: @:| and the ^horrors of :living in :this 'part
of the c/ollege#
or ^anything like th\at#
" - laugh)"

KEY

_ are features of stress, including boosters
\ are tones (fall, rise, and fall-rise)
– are pauses
≠ is a tone unit boundary
@: is a pause filler *'er'*
( ) enclose a subordinate tone unit
" " enclose contextual comments
( ) enclose uncertain material

شکل ۱-۳ تقطیع پیرایشی (Skeleton parsing)، از یک پیکرهٔ گفتاری زبان شفاهی قلاب‌های

بی‌نام نشان دهندهٔ سازه‌ای هستند که در طرح نشانه‌گذاری برای آنها رجوعی ارائه نشده است. (نشانه‌گذارها مجاز هستند که بدون نشان دادن جزییات سازه‌ها به آنهاتعلق دارند، آنها را مشخص نمایند.) (لیچ، ۱۹۹۳: ۲۷۸)

[S[N theme_NP1 _ [N the_AT killer_NN1 whale_NN1 N] _ [Fr[N
who_PNQS \[V 'd_VHD grown_VVN [J too_RG big_JJ [P for_IF
[N its_APPS pool_NN1 [P on_II [N Clacton_NP1 Pier_NN1]
N]P\P]J]\F-[N] _ [V has_VHZ arrived_VVN safely_RR [P at_II
[N its_APPS new_JJ home_NN1 [P in_II [N Windsor_NP1] safari_NN1
park_VNL\ N]P]N]F[V] _ S]

KEY: Square brackets enclose constituents above word level. Headings are linked to their words by '-'. The tagset used here is the revised version of the LOB tagset, known as the 'CLAWS2 tagset'. Tag definitions can be inferred from a comparison between this example and Fig. 2. Symbols for constituents above word level are: F, relative clause; J, adjective phrase; N, noun phrase; P, prepositional phrase; S, sentence; V, verb phrase

شکل ۱-۴ تقطیع [illegible] (۱۹۹۳: ۲۷۸)

```
[S [illegible]
        [NP Mr [illegible]]
        [VP is
            [NP chairman
                [PP of
                    [NP [NP Elsevier [illegible])
                        [NP the Dutch
                            publishing
                            group) [illegible] ])
]
```

شکل ۱-۵ نوعی برچسب [illegible] (۱۹۹۳: ۲۱۷)

[illegible]	The	Z8
[illegible]	[illegible]	E4.1+
[illegible]	[illegible]	O4.2-
[illegible]	[illegible]	A1.1.1
APP[illegible]	[illegible]	Z8
NN[illegible]	[illegible]	F3
[illegible]	[illegible]	Z5
[illegible]	the	Z5
NN[illegible]	[illegible]	F1
CC	and	Z5
[illegible]	[illegible]	M1
[illegible]	[illegible]	Z5
[illegible]	the	Z5
NN[illegible]	[illegible]	Q1.3
TO	[illegible]	Z5
[illegible]	[illegible]	Q1.3[illegible]2.1
[illegible]	[illegible]	Q1.3[illegible]2.2
DB	all	N5.1+
APP[illegible]	[illegible]	Z8
NN[illegible]	[illegible]	S2/S3[illegible]

KEY: A1[illegible] general actions; E4.1, happy and sad; F1, food; F3, cigarettes and drugs; [illegible] motion; N5.1, all, none; O4.2, judgement of appearance [illegible]; Q1.3, telecommunications; S2, people; S3.1, relationships (general); Z5, function words; Z8, pronouns, etc.

شکل ۱-۶ نشانه‌گذاری کلامی ... رجوعی ... سوشیتدپرس (Associated Press)

(لیچ، ۱۹۹۳: ۲۰۹)

S.1 (0) The state Supreme Court has refused to release [1 [2 Rahway State Prison 2] inmate 1] [1 James Scott 1] on bail.

S.2 (1 The fighter 1] is serving 30-40 years for a 1983 armed robbery conviction.

S.3 (1 Scott 1] had asked for freedom while <1 he awaits for an appeal decision.

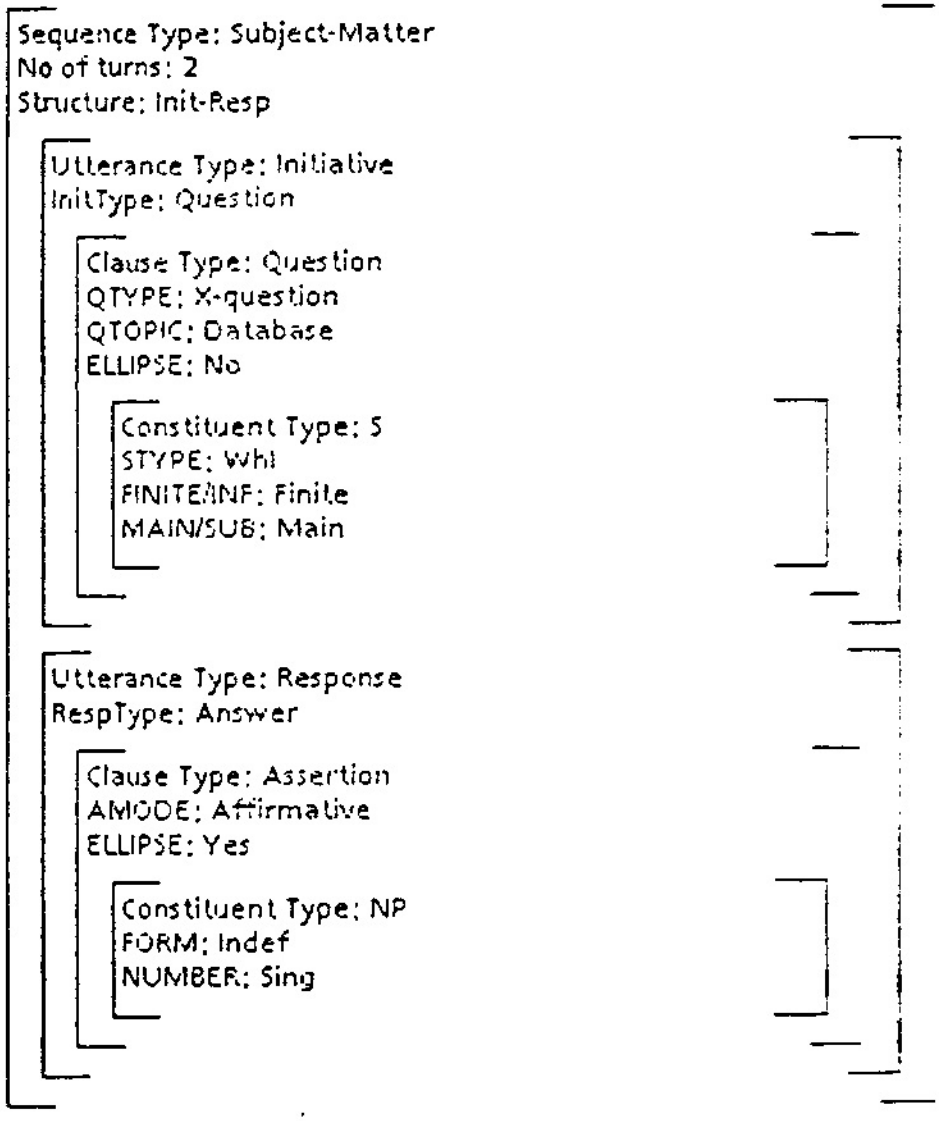
S.4 Meanwhile, [2 <1 his promoter 3] [3 Murad Muhammed 3], said Wednesday <3 he settled only $[[2,250 or 4 [1 Scott 1] is nationally televised light heavyweight fight against 5 ranking contender 5]] (5 Yaqui Lopez 5] last Saturday 4).

S.5 (4 The fight in which 1 Scott 1] won a unanimous decision over (5 Lopez 5 4] grossed $[[35,000 for [3 Muhammed 3]'s firm 6], ((6 Triange Productions in Newark 6]), he said.

KEY: The use of the same index *n* binds one syntactic constituent to another to which it is co-referential or semantically equivalent. In the following list, *n* represents an arbitrary index:

(i i) *or* [i i]	enclose a constituent (normally a noun phrase) entering into an equivalence chain
<i	indicates a pronoun with a preceding antecedent
>i	indicates a pronoun with a following antecedent
{{i i}	enclose a noun phrase entering into a copular relationship with a preceding noun phrase
{i i}}	enclose a noun phrase entering into a copular relationship with a following noun phrase
(0)	represents an anaphoric barrier; in effect, the beginning of a new text

شکل ۷- بخشی از ساخت برچسبی برای یک مکالمه (آهرن برگ و یونسن، ۱۹۸۸: ۳۱)



۱ـ۲ـ۷ نشانه‌گذاری کارکرد شناختی /کلامی

اصطلاح پوششی «نشانه‌گذاری کلامی» را می‌توان برای هر نوع نشانه‌گذاری که به واحدها یا روابطی ورای مرز جمله برمی‌گردد، به‌کار برد. برای داده‌هایی از زبان گفتاری نوعی نشانه‌گذاری به کار رفته است (اشتن اشتروم[۲۹]، ۱۹۸۴) تا طبقاتی از علائم کلامی را که نشان‌دهندهٔ ساختار مکالمه می‌باشند نشان دهد. نوع دیگری از نشانه‌گذاری که تا زمان نگارش مقالهٔ (لیچ، ۱۹۹۳) تنها روی داده‌های نوشتاری در دانشگاه لنکستر به کار گرفته شده بود، نشان دهندهٔ *روابط ارجاعی*[۳۰] و روابط همسنگی[۳۱] دیگر میان اجزاء همسایه در یک متن است. نمونه‌ای از متنی که به این روش نشانه‌گذاری شده در شکل ۱ـ۶ آمده است. آهرن برگ و یونسن[۳۲] (۱۹۸۸: ۷۰ـ۶۶) از دانشگاه لینکوپینگ[۳۳] سوئد نیز یک سیستم محاوره‌ای[۳۴] را برای برچسب دهی مکالمات معرفی می‌کنند که در آن برچسب‌ها برای نشان دادن اطلاعات کارکرد شناختی به کار می‌روند. این اطلاعات برای مثال عبارتند از: نشان دادن برنهاده[۳۵](ها) در یک نقطهٔ خاص ازکلام یا کنش(های) گفتاری که در یک پاره‌گفتار ویژه اجرا می‌شود. شکل ۱ـ۷ نمونه‌ای از ساختار برچسب‌های این سیستم را که Dagtag نام دارد، نشان می‌دهد.

۱ـ۳ چکیده

در این فصل شاخهٔ میان رشته‌ای زبانشناسی پیکره بنیاد به طور اجمالی معرفی گردید و ضمن ذکر مراحل ایجاد یک پیکرهٔ رایانه‌ای، مسألهٔ نشانه‌گذاری پیکره نیز به میان آورده شد. گفته شد که نشانه‌گذاری پیکره عبارت است از افزودن اطلاعات تفسیری عمدتاً زبانی به یک پیکرهٔ گفتاری و یا نوشتاری زبان با استفاده از نوعی رمزگذاری بر روی نمود الکترونیک مادهٔ زبانی. سپس انواع نشانه‌گذاری پیکره به اختصار معرفی گردید. فصل بعدی اختصاص دارد به معرفی برچسب‌دهی دستوری پیکره.

---

29. Stentröm
30. anaphoric relations
31. equivalence relations
32. L.Ahrenberg and A.Jönsson
33. Linköping
34. interactive
35. topic

فصل ۲ برچسب دهی دستوری

برچسب دهی دستوری، یا مشخص نمودن مقولهٔ دستوری هر واژه، نوعی نشانه‌گذاری است که بیش از همه در پیکره‌های تهیه شده در اروپا و آمریکا رایج است و این خود دو دلیل دارد:

۱ـ این کار به اندازه‌ای ساده هست که بتوان آن را تا حد زیادی به صورت خودکار انجام داد.

۲ـ کاربردهای فراوانی دارد، مثلاً یکی در فرهنگ‌نگاری که در آن برچسب دهی دستوری قدم اول است در عمل سرواژه‌یابی[1] (لیچ، ۱۹۹۳: ۲۷۶). ضمناً همان گونه که در بخش ۱ـ۲ـ۵ ذکر شد، این کار اولین گام در تقطیع نحوی جملات نیز هست.

یک طرح برچسب دهی دستوری شامل موارد زیر است (لیچ، ۱۹۹۳:۲۷۶):

۱ـ یک مجموعه برچسب[2]: مجموعه‌ای از عنوان‌های دستوری برای اقلام واژگانی

۲ـ مجموعه‌ای از تعاریف برچسب‌ها

۳ـ مجموعه‌ای از دستورالعمل‌های برچسب‌دهی که نحوهٔ افزودن برچسب‌ها را به متن مشخص می‌کند.

هنگامی که اقلام واژگانی تحلیل شدند و مقولهٔ دستوری هر یک مشخص شد، برچسب مربوط به مقولهٔ دستوری هر واژه -مانند هر نوع عمل نشانه‌گذاری دیگر (ر.ک. ۱ـ۲)- در کنار آن واژه وارد متن می‌گردد. البته این کار باید به طریقی انجام شود که باز بتوان به متن بدون برچسب‌ها را حذف کرد و به متن اصلی دست یافت. شکل ۲ـ۱ نمونه‌ای از یک متن برچسب داده شده به همراه معانی برچسب‌های آن را نشان می‌دهد.

۲ـ۱ تاریخچهٔ نشانه‌گذاری نحوی

عمل برچسب دهی دستوری یک پیکرهٔ زبانی تاکنون در ایران انجام نپذیرفته است. حتی تهیهٔ یک پیکره نیز از سابقه‌ای طولانی درکشور ما برخوردار نیست، بلکه از اوایل

---

1. lemmatization 2. tagset

شکل ۲-۱ برچسب‌دهی دستوری، پیکرهٔ LOB توجه کنید که در این طرح، علائم طرح‌های دیگر برای متون نوشتاری، علائم نقطه‌گذاری (و غیره) به حساب آمده و برچسب‌دهی می‌شوند.

[illegible]_NN is_BEZ an_AT excellent_JJ virtue_NN ,_
but_CC not_XNOT when_WRB the_ATI guests_NNS have_HV
to_TO sweep_VB in_IN rows_NNS in_IN the_ATI cellar_NN
-

the_ATI covers_NNS ,_ whose_WP$ chief_JJB scene_NN
was_BEDZ cut_VBN at_IN the_ATI last_AP moment_NN ,_
had_HVD comparatively_RB little_AP to_TO sing_VB

_ he_PP3A stole_VBD my_PP$ wallet_NN !_!_!
named_VBD Rollinson_NP ._

(Key: AP, post-determiner; AT, singular article; ATI, article neutral for number; BEDZ, *was* (past sing. form of the verb *be*); BEZ, *is* (-s form of the verb *be*); CC, coordinating conjunction; HV, base form of the verb *have*; HVD, *had*, *'d* (past form of the verb *have*); IN, preposition; JJ, general adjective; JJB, attributive-only adjective; NN, singular common noun; NNS, plural common noun; NP, singular proper noun; PP3A, third-person singular personal pronoun, subjective; PP$, possessive pronoun (determinative function); RB, general adverb; TO, *to*-infinitive marker; VB, base form of lexical verb; VBD, past tense of lexical verb; VBN, past participle of lexical verb; WP$, possessive *wh*-pronoun (*whose*); XNOT, negative particle (*not*, *n't*)