

به نام خداوند مهربان



وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

مجری

الهام علایی ابودر

مهر ۹۸

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی «امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» پیرو قرارداد شماره ۳۳/۴۵۹۹ مورخ ۱۳۹۷/۴/۲۳ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و خانم دکتر الهام علایی ابوذرا اجرا شده است. گزارش حاضر مستند این طرح است. این گزارش و تمامی حقوق مادی آن بر اساس «قانون حمایت از حقوق مؤلفان و مصنفان و هنرمندان» مصوب سال ۱۳۸۴ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون، متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران است و هرگونه استفاده از تمام یا پاره‌ای از آن شامل نقل‌قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آن به‌صورت چاپی، الکترونیکی یا وسایل دیگر تنها با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل‌قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.



بناام خدا

آغاز نیم قرن دوم خدمات ارزشمند ایرانداک به علم، فناوری، و نوآوری گرامی باد

۱۳۴۷ - ۱۳۹۸

گواهی

انجام طرح پژوهشی:

◇ عنوان: امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

◇ مجری: دکتر الهام علایی ابوذر

◇ همکار(ان): -

◇ شماره قرارداد: ۳۳/۴۵۹۹

◇ تاریخ قرارداد: ۱۳۹۷/۴/۲۳

◇ تاریخ شروع: ۱۳۹۷/۴/۲۳

◇ تاریخ پایان: ۱۳۹۸/۶/۳۰

◇ ناظر: دکتر سیروس علی‌دوستی

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایه داوری در نشست ۳۴۵ (تاریخ ۱۳۹۸/۷/۹) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری و ناظر این طرح تقدیم دارم.

با آرزوی بهروزی
پرویز شهریاری
معاون پژوهش و آموزش

**امكان سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات
ایران (ایرانداک)**

مدیر طرح پژوهشی: الهام علایی ابوزر، عضو هیات علمی پژوهشکده علوم اطلاعات پژوهشگاه علوم و فناوری
اطلاعات ایران (ایرانداک)

نشانی: تهران، خیابان انقلاب، تقاطع فلسطین، شماره ۱۰۹۰

تلفن: ۶۶۹۵۱۴۳۰ (۰۲۱) **دورنگار:** ۶۶۴۶۲۲۵۴ (۰۲۱)

رایانامه: alayi@irandoc.a.ir

چکیده

پیکره‌های زبانی و واکاوی پیکره‌بنیاد زبان در زبان‌شناسی پیکره‌ای نقشی کلیدی دارند. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون، و میان‌رشته است. بنابراین برای کارایی و اثربخشی در ساخت آن‌ها نیاز است که پیش از آغاز کار، امکان‌سنجی انجام شود. امکان‌سنجی؛ گام‌ها، هزینه‌ها، نیروی انسانی، حقوق مادی و معنوی، و مانند آن‌ها را برای یک پروژه، بررسی و به مدیریت پروژه کمک می‌کند تا آن را با آمادگی و آینده‌نگری بیشتری به سرانجام برساند. از آنجایی که مدلی برای امکان‌سنجی ساخت پیکره متنی همچون گونه‌ای از پیکره‌ها در پژوهش‌های پیشین نبود، این پژوهش در پی ساخت چنین مدلی انجام شد. برای این کار، پس از بررسی پژوهش‌های پیشین، شاخص‌های کلی ساخت پیکره‌های زبانی استخراج شد. سپس، فرآیند ساخت پیکره به تفصیل مورد بررسی قرار گرفت. پس از بررسی پژوهش‌های پیشین در زمینه امکان‌سنجی، بر پایه اطلاعات به دست آمده از مطالعات امکان‌سنجی، بررسی شاخص‌های ساخت پیکره و فرآیند ساخت پیکره، ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره استخراج شد و مدلی کلی برای ساخت پیکره پیشنهاد شد. سپس با روش دلفی و در دو دور اعتباریابی، مدل پایانی به دست آمد. این مدل دارای هفت بُعد فنی، اقتصادی، زمان‌بندی، قانونی حقوقی، عملیاتی، تأمین مالی، و بازاریابی و روی هم رفته ۳۳ مؤلفه است. این مدل و کاربرد آن در پروژه‌های ساخت پیکره متنی می‌تواند کمک بزرگی برای پژوهشگران و سازمان‌ها باشد. سپس در آخرین مرحله از پژوهش، مدل امکان-سنجی ساخت پیکره در پنلی متشکل از متخصصان موضوعی و اجرایی در ایرانداک به شور گذاشته و اعتباریابی شد و مدل پایانی برای ساخت پیکره از داده‌های علمی ایرانداک به دست آمد و در پایان، امکان‌سنجی ساخت پیکره از داده‌های ایرانداک نیز انجام شد و چگونگی پرداختن به ابعاد و مولفه‌های مدل مشخص شد.

واژه‌های کلیدی: زبان‌شناسی پیکره‌ای؛ پیکره زبانی، امکان‌سنجی؛ مدل؛ روش دلفی

فهرست مطالب

عنوان	شماره صفحه
فصل اول:	
مقدمه.....	۱
۱-۱- مقدمه.....	۲
۱-۲- پیکره زبانی.....	۲
۱-۳- زبان‌شناسی پیکره‌ای.....	۵
۱-۴- کاربرد پیکره در مطالعات زبان‌شناسی و پردازش زبان.....	۱۰
۱-۵- معرفی برخی از پیکره‌های زبانی فارسی.....	۱۱
۱-۶- ضرورت انجام پژوهش.....	۱۷
۱-۷- سوال پژوهش.....	۱۸
فصل دوم: پیشینه پژوهش.....	۲۰
۲-۱- مقدمه.....	۲۱
۲-۲- مرور پژوهش‌های پیشین ساخت پیکره زبانی.....	۲۱
۲-۳- خلاصه فصل.....	۲۷
فصل سوم: مبانی نظری ساخت پیکره زبانی.....	۲۸
۳-۱- مقدمه.....	۲۹

۲۹.....	۲-۳- شاخص‌های کلی ساخت پیکره زبانی.....
۳۰.....	۳-۲-۱- نمونه‌گیری.....
۳۵.....	۳-۲-۲- نمایندگی.....
۳۷.....	۳-۲-۳- توازن.....
۳۹.....	۳-۲-۴- اندازه.....
۴۰.....	۳-۲-۵- نوع پیکره (عمومی یا تخصصی).....
۴۱.....	۳-۲-۶- یک‌دستی.....
۴۲.....	۳-۳- فرآیند ساخت پیکره متنی.....
۴۳.....	۳-۱-۳- پیش‌پردازش متون.....
۴۶.....	۳-۲-۲- حاشیه‌نویسی.....
۵۳.....	۳-۴-۲- فرآیند ساخت پیکره گفتاری.....
۵۵.....	۳-۴-۱- جمع‌آوری داده‌ها.....
۵۵.....	۳-۴-۲- آوانویسی.....
۵۶.....	۳-۴-۳- نمایش و حاشیه‌نویسی.....
۵۷.....	۳-۴-۴- دسترسی.....
۵۹.....	۳-۵- خلاصه فصل.....
۶۰.....	فصل چهارم: توسعه چارچوب مفهومی امکان‌سنجی ساخت پیکره زبانی.....
۶۱.....	۴-۱- مقدمه.....
۶۱.....	۴-۲- تعریف امکان‌سنجی.....
۶۲.....	۴-۳- بررسی پژوهش‌های امکان‌سنجی.....
۶۴.....	۴-۴- ابعاد امکان‌سنجی.....

۶۴.....	۱-۴-۴- امکان‌سنجی فنی.....
۶۶.....	۲-۴-۴- امکان‌سنجی زمان‌بندی.....
۶۷.....	۳-۴-۴- امکان‌سنجی اقتصادی.....
۶۸.....	۴-۴-۴- امکان‌سنجی عملیاتی.....
۶۹.....	۵-۴-۴- امکان‌سنجی عوامل انسانی.....
۷۰.....	۶-۴-۴- امکان‌سنجی قانونی/حقوقی.....
۷۱.....	۷-۴-۴- امکان‌سنجی فرهنگی.....
۷۵.....	۵-۴- اعتباریابی چارچوب مفهومی امکان‌سنجی ساخت پیکره زبانی.....
۷۷.....	۱-۵-۴- فرآیند ساخت پرسش‌نامه.....
۹۶.....	۲-۵-۴- دور اول دلفی.....
۱۰۲.....	۳-۵-۴- دور دوم دلفی.....
۱۰۷.....	۶-۴- مدل امکان‌سنجی ساخت پیکره زبانی (پیکره متنی).....
۱۱۱.....	۷-۴- خلاصه فصل.....
۱۱۲.....	فصل پنجم: امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک.....
۱۱۳.....	۱-۵- مقدمه.....
۱۱۳.....	۲-۵- طراحی مدل امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک.....
۱۱۸.....	۳-۵- اعتباریابی مدل امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک.....
۱۲۵.....	۴-۵- امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک.....
۱۳۶.....	۵-۵- خلاصه فصل.....
۱۳۷.....	فصل ششم: نتیجه‌گیری و پیشنهاد برنامه ساخت پیکره زبانی از داده‌های علمی ایرانداک.....
۱۳۸.....	۱-۶- بحث و نتیجه‌گیری.....

۶-۲- مقایسه یافته‌ها با پژوهش‌های پیشین..... ۱۴۲

۶-۳- پیشنهاد برنامه ساخت پیکره زبانی از داده‌های علمی ایرانداک..... ۱۴۲

فهرست منابع فارسی..... ۱۵۲

فهرست منابع لاتین..... ۱۵۵

فهرست جدول‌ها

جدول ۳-۱: مشخصات متون پیکره بی جن خان (برگرفته از بی جن خان و همکاران: ۲۰۱۱).....	۳۴
جدول ۳-۲: ساختار پیکره ملی بریتانیایی: نوع متنی این پیکره (برگرفته از شاروف: ۲۰۰۳، نقل از مک‌انری و همکاران: ۲۰۰۶).....	۳۹
جدول ۳-۳: حوزه نوشتاری متون (برنارد: ۲۰۰۷).....	۵۲
جدول ۳-۴: تاریخ انتشار متون (برنارد: ۲۰۰۷).....	۵۲
جدول ۳-۵: وسیله بیان مربوط به متون نوشتاری (برنارد: ۲۰۰۷).....	۵۲
جدول ۳-۶: چهار رویکرد نمایش نمادین وقفه و طول وقفه (برگرفته از جانسون: ۱۹۹۵، نقل از تامسون: ۲۰۰۴).....	۵۶
جدول ۴-۱: ابعاد امکان‌سنجی و مولفه‌های مربوط به هر یک از ابعاد.....	۷۹
جدول ۴-۲: پرسش‌نامه اولیه ساخت پیکره متنی.....	۸۱
جدول ۴-۳: نحوه اظهار نظر متخصصان درباره کلیت پرسش‌نامه امکان‌سنجی ساخت پیکره متنی.....	۸۵
جدول ۴-۴: پاسخ متخصصان زبان‌شناسی رایانشی درباره روایی پرسش‌نامه اولیه ساخت پیکره متنی.....	۸۶
جدول ۴-۵: روش اعمال یا رد نظر متخصصان رایانشی (همراه با ارائه دلیل عدم پذیرش برخی از نظرات) درباره روایی پرسش‌نامه اولیه ساخت پیکره متنی.....	۸۹
جدول ۴-۶: پرسش‌نامه ساخت پیکره متنی، پس از ویرایش توسط متخصصان.....	۹۲
جدول ۴-۷: درصد فراوانی انتخاب گزینه‌های مربوط به هر گویه در بخش نخست دور اول دلفی.....	۹۶
جدول ۴-۸: میانگین پاسخ به هر گویه در دور اول دلفی.....	۹۸

جدول ۴-۹: نظر پاسخ‌گویان درباره سایر مولفه‌های احتمالی دیگر در ساخت پیکره.....	۱۰۰
جدول ۴-۱۰: فهرست سایر ابزارهای پردازش زبان طبیعی و انواع حاشیه‌نویسی.....	۱۰۱
جدول ۴-۱۱: روش اعمال یا رد نظر متخصصان درباره دیدگاه‌های مطرح شده در پرسش‌نامه دور اول دلفی.....	۱۰۲
جدول ۴-۱۲: پرسش‌نامه دور دوم دلفی (به مشورت گذاردن نظرات متخصصان).....	۱۰۴
جدول ۴-۱۳: درصد فراوانی پاسخ به هر گویه در پرسش‌نامه دور دوم دلفی.....	۱۰۶
جدول ۴-۱۴: میانگین پاسخ به هر گویه در پرسش‌نامه دور دوم دلفی.....	۱۰۶
جدول ۴-۱۵: مدل امکان‌سنجی ساخت پیکره متنی.....	۱۰۸
جدول ۵-۱: پرسش‌نامه ساخت پیکره متنی از داده‌های علمی ایرانداک.....	۱۱۴
جدول ۵-۲: درصد فراوانی پاسخ به هر گویه در پرسش‌نامه ساخت پیکره متنی از داده‌های علمی ایرانداک.....	۱۱۹
جدول ۵-۳: میانگین پاسخ به هر گویه در امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک.....	۱۲۰
جدول ۵-۴: مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک.....	۱۲۳
جدول ۵-۵: مشخص نمودن چگونگی پرداختن به هر یک از ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک.....	۱۲۷

فهرست شکل ها

- شکل ۳-۱: شاخص های اصلی تهیه پیکره زبان..... ۴۲
- شکل ۳-۲: فرآیند ساخت پیکره متنی..... ۵۱
- شکل ۳-۳: فرآیند ساخت پیکره گفتاری..... ۵۸
- شکل ۴-۱: مدل امکان سنجی ارائه محصولات بیمه ای موجود به صورت الکترونیک در برخی از شرکت های بیمه (صرافی - زاده و علیپور: ۱۳۹۰) ۷۲
- شکل ۴-۲: ابعاد امکان سنجی..... ۷۴
- نمودار ۴-۳: میانگین پاسخ به هر گویه در دور اول دلفی..... ۹۹
- نمودار ۴-۵: میانگین پاسخ به هر گویه در پرسش نامه دور دوم دلفی..... ۱۰۷
- شکل ۴-۶: مدل امکان سنجی ساخت پیکره متنی..... ۱۰۹
- نمودار ۵-۱: میانگین پاسخ به هر گویه در امکان سنجی ساخت پیکره از داده های علمی ایرانداک..... ۱۲۱
- شکل ۵-۲: مدل امکان سنجی ساخت پیکره متنی از داده های علمی ایرانداک..... ۱۲۴

سپاسگزاری

از ناظر محترم، جناب آقای دکتر علیدوستی، برای همراهی گام به گام این پژوهش، کمال تشکر را می‌نمایم. بسیار از رهنمودهای ارزشمند ایشان در طراحی و اعتباریابی مدل امکان‌سنجی ساخت پیکره زبانی و مدل امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک، بهره بردم.

فصل اول

مقدمه

۱-۱- مقدمه

طی سال‌های اخیر تهیه پیکره‌های زبانی و تجزیه و تحلیل پیکره-بنیاد زبان بسیار مورد توجه پژوهشگران قرار گرفته است. بسیاری از پژوهش‌های زبان‌شناختی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکره زبانی امکان‌پذیر است؛ در این میان تهیه پیکره‌های متنی که حاوی نمونه‌های خوب و مناسب باشد از جایگاه خاص برخوردار است. در این بخش به تعریف پیکره زبانی، معرفی حوزه زبان‌شناسی پیکره‌ای، کاربرد پیکره در مطالعات زبان‌شناسی و پردازش زبان، معرفی برخی از پیکره‌های زبان فارسی، ضرورت انجام پژوهش حاضر، سوال و فرضیه/فرضیه‌های پژوهش پرداخته می‌شود.

۱-۲- پیکره زبانی

برای تعریف پیکره، برای پیشگیری از هر گونه ابهام، ابتدا بهتر است بدانیم چه چیزی عمدتاً پیکره محسوب نمی‌شود؛ مجموعه متون زبانی بسیاری وجود دارد که نمی‌توان آن‌ها را پیکره نامید. به عنوان مثال، وب جهانی پیکره نیست، زیرا مدام در حال تغییر است و بر اساس رویکردهای زبان‌شناسی طراحی نشده است. آرشیو نیز پیکره محسوب نمی‌شود، زیرا دلایل جمع‌آوری متون در پیکره و آرشیو متفاوت است و هر کدام اولویت‌های خاصی را در این راستا دنبال می‌کنند (سینکلر: ۲۰۰۴). پیکره^۱، مجموعه‌ای نظام‌مند، رایانه‌ای شده، معتبر و موثق از زبان است که جهت مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد. مجموعه داده‌های زبانی موجود در پیکره هم می‌تواند به صورت متن و هم برگردان مکتوب گفتار ضبط شده باشد. اصولاً هر مجموعه‌ای که متشکل از بیش از یک متن باشد را می‌توان پیکره متنی^۲ خواند؛ اما واژه «پیکره» یا عبارت «پیکره متنی» نوعاً در زبان‌شناسی مدرن به کار می‌رود و به مجموعه‌ای از متون اطلاق می‌شود که جهت تجزیه و تحلیل زبان‌شناسی یا پردازش زبان طبیعی مورد استفاده قرار می‌گیرد. پیکره می‌تواند هم از متون کامل تشکیل شده باشد و هم از گزیده‌هایی بسیار طولانی از متون کامل.

¹ corpus

² text corpus

معمولاً متون پیکره‌ها به نحوی انتخاب می‌شوند که نماینده نوعی خاص از زبان یا گونه‌ای از زبان باشند؛ مثلاً ممکن است پیکره به شکلی جمع‌آوری شود که زبان انگلیسی به کار رفته در کتاب‌های تاریخی یا گونه کانادایی زبان فرانسوی را نمایندگی کند. در تهیه پیکره متنی، ویژگی‌هایی مانند: نمونه‌گیری و نمایندگی، اندازه، قابلیت خوانش توسط رایانه و ارجاع استاندارد و معتبر، لحاظ می‌شود. از آنجایی که تحلیل داده‌ها در پیکره مورد مطالعه، بر اساس شمارش دقیق جمله‌ها، عبارات، واژه‌ها و حتی تکواژهاست، متون مورد مطالعه حتماً باید قابلیت خوانش توسط رایانه را داشته باشند. این متون معمولاً به شکل مجموعه داده‌های الکترونیکی ذخیره می‌شوند. در حقیقت، پیکره مجموعه‌ای نسبتاً بزرگ از متون الکترونیکی است که بصورت حساب‌شده‌ای حاشیه‌نویسی، برچسب‌گذاری و دسته‌بندی شده‌اند و در نتیجه امکان بررسی‌های بسیار متنوعی را برای کاربر فراهم می‌آورند (مک انری و ویلسون: ۲۰۰۱). اتکینز و همکاران (۱۹۹۲، نقل از کوهستانی: ۱۳۸۹) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و برای آن ۹ تقسیم‌بندی انجام داده‌اند که عبارتند از:

- تقسیم‌بندی به انواع «متن کامل»^۱، «نمونه‌ای»^۲ و «نظارتی»^۳. منظور از پیکره متن کامل، پیکره‌ای است که در آن همه متون، کامل هستند؛ به عبارتی، متون خلاصه نشده‌اند. در پیکره‌های نمونه‌ای، حجم نمونه و جایگاه آن در متن اصلی، از پیش تعریف می‌شود. در پیکره‌های نظارتی، متون به طور مداوم بازبینی و برای گنجانده شدن در پیکره، فیلتر می‌شوند.
- تقسیم‌بندی به انواع «بسته»^۴ و «باز»^۵. در پیکره‌های باز امکان افزودن داده در مراحل بعدی وجود دارد؛ درحالی‌که این امکان در مورد پیکره‌های بسته، وجود ندارد.
- تقسیم‌بندی به انواع «هم‌زمانی»^۶ و «درزمانی»^۷. پیکره‌های هم‌زمانی حاوی داده‌های معاصر زبانی هستند و تحولات تاریخی و سیر پیشرفت زبان مورد مطالعه را در نظر نمی‌گیرند. پیکره‌های درزمانی، پیکره‌هایی هستند که حاوی متونی از دوره‌های گوناگون هستند و هدف آن‌ها بررسی تغییرات زبانی در دوره‌های مدنظر است.

¹ whole text

² samples

³ monitor

⁴ open

⁵ closed

⁶ synchronic

⁷ diachronic

- تقسیم‌بندی به انواع «عمومی^۱» و «اصطلاح‌شناسی^۲». در پیکره‌های اصطلاح‌شناسی، اصطلاح‌شناسان شرایطی را تعیین می‌کنند که پیکره در مصارف اصطلاح‌شناسی به کار رود.
- تقسیم‌بندی به انواع «یک‌زبان^۳»، «دو‌زبان^۴» و «چندزبان^۵».
- تقسیم‌بندی بر اساس زبان به کار رفته در پیکره (مثلاً فارسی، انگلیسی، روسی و ...).
- تقسیم‌بندی به انواع «منفرد^۶»، «موازی^۷»: پیکره منفرد از متونی تشکیل می‌شود که به یک زبان واحد هستند و با متونی از زبان/ زبان‌های دیگر در ارتباط نیستند. اما پیکره‌های موازی از متونی تشکیل شده‌اند که با ترجمه‌هایی از یک یا بیش از یک زبان دیگر در ارتباط قرار گرفته‌اند.
- تقسیم‌بندی به انواع «مرکزی^۸» و «پوسته‌ای^۹». پیکره مرکزی، مجموعه‌ای منتخب از متون است که حجمی قابل مدیریت دارد و برای کاربردهای معمول، مفید است. در مقابل، پیکره پوسته‌ای از باقی‌مانده کتابخانه متون الکترونیکی، تشکیل شده است و تنها در زمان نیاز به کار می‌رود.
- تقسیم‌بندی به انواع «هسته‌ای^{۱۰}» و «پیرامونی^{۱۱}». نوع هسته‌ای شامل انواع متونی است که در همه گونه‌های زبانی مشترک هستند و بنابراین در تمام پیکره‌های فرعی موجودند، اما پیکره‌های پیرامونی شامل متونی است که تنها در برخی از پیکره‌های فرعی وجود دارند.

البته تقسیم‌بندی‌های دیگری نیز بر اساس شکل داده‌ها (پرینت‌شده، متن الکترونیکی، گفتار دیجیتالی - شده، ویدئویی یا تلفیقی از چند شکل)، روش طراحی پیکره (متوازن، هرمی، ...)، اندازه پیکره (اندازه ثابت، اندازه کنترل‌شده)، استفاده از متن اصلی یا ترجمه‌شده، نوع استفاده (عمومی یا تخصصی)، گفتاری^{۱۲} یا نوشتاری بودن پیکره، نوع حاشیه‌نویسی (ساده، نشانه‌گذاری شده) انجام شده است.

¹ general
² thesaurus
³ monolingual
⁴ bilingual
⁵ multilingual
⁶ single
⁷ parallel
⁸ central
⁹ cluster
¹⁰ nuclear
¹¹ Perimeter
¹² spoken corpus

متون موجود در پیکره‌های متنی حاوی هزاران/ میلیون‌ها کلمه است که گویشوران در بافت طبیعی گفتار یا نوشتار استفاده می‌کنند. در تهیه پیکره‌های متنی از منابع گوناگونی استفاده می‌شود که مهم‌ترین آن‌ها متون روزنامه‌ها و کتاب‌ها است. امروزه صفحات وب نیز به عنوان منبع غنی تهیه پیکره‌های متنی مورد استفاده قرار می‌گیرند؛ زیرا صفحات وب امکان جمع‌آوری متون مختلف، به نحوی که نماینده زبان مورد مطالعه باشد، را فراهم می‌آورند و این امر از طریق جمع‌آوری متون گوناگون مربوط به ژانرها و نویسندگانه‌های گوناگون انجام می‌پذیرد. در نتیجه پیکره متنی می‌تواند تصویر معقولی از زبان مورد مطالعه ارائه دهد.

۱-۳- زبان‌شناسی پیکره‌ای

پیکره‌های زبانی نوعاً در پردازش زبان طبیعی و زبان‌شناسی پیکره‌ای^۱ مورد استفاده قرار می‌گیرند. گرچه در گذشته، بسیاری از زبان‌شناسان بر اهمیت پیکره زبانی در بیشتر بررسی‌هایشان تاکید کرده‌اند، اما در دوران جدید است که تکیه بر داده‌های واقعی زبان به صورت گسترده‌ای رواج یافته و شرط اساسی بسیاری از پژوهش‌های نظری و کاربردی مانند: نظریه‌پردازی و توصیف ساختمان زبان، گویش‌شناسی، دستورنویسی و فرهنگ‌نگاری به شمار می‌آید. امکان سازمان‌دهی، تنظیم، تفکیک، جستجو و دستیابی سریع داده‌های زبانی، افق‌های تازه‌ای در برابر پژوهندگان گشوده و باعث پیدایش شاخه‌ای تخصصی در حوزه زبان‌شناسی رایانه‌ای گردیده است. این شاخه با نام زبان‌شناسی پیکره‌ای، تنها در آخرین دهه‌های قرن بیستم ایجاد شده است و در همین زمان کوتاه تبدیل به یکی از فعال‌ترین و پرکاربردترین زمینه‌ها شده است. تالیف کتاب‌های مختلف و ایجاد مجامع علمی، مراکز پژوهشی، نشریه‌های تخصصی و بین‌المللی و سایت‌های اینترنت برای این شاخه، گویای این واقعیت است (عاصی: ۱۳۸۲). زبان‌شناسی پیکره‌ای به مطالعه فقره‌های واژگانی، ساخت‌های دستوری یا ارتباط این دو با سایر مشخصه‌های زبانی و غیرزبانی می‌پردازد. زبان‌شناسی پیکره‌ای را می‌توان گذار از رویکردهای ذهن‌گرایانه به پژوهش‌های زبانی دانست. در رویکردهای ذهن‌گرایانه بر کشف فرآیندهای در حال وقوع در ذهن بشر تمرکز می‌شود. در مقابل، تحقیقات پیکره‌ای بر الگوهای واقعی کاربرد زبان در پایگاه‌های حجیم دادگان یا همان پیکره‌ها تمرکز دارد. زبان‌شناسی پیکره‌ای هم از روش‌های کمی و هم از روش‌های کیفی برای تحلیل زبان بهره می‌برد و

¹ corpus linguistics

رایانه را به منظور انجام تحلیل‌های پیچیده به کار می‌گیرد. آنچه پیش از پرداختن به زبان‌شناسی پیکره‌ای باید مدنظر قرار گیرد، این است که این زیرشاخه از زبان‌شناسی را باید رویکردی مکمل برای رویکردهای سنتی‌تر به بررسی زبان در نظر گرفت. در واقع آنچه اصولاً باعث می‌شود زبان‌شناسی پیکره‌ای وارد کار زبان گردد، قوت آن در مطالعه کاربرد واقعی زبان است تا پژوهشگر از افتادن به ورطه نظریه‌پردازی صرف، مصون بماند و باز همین نقطه قوت است که تحلیل پایگاه‌های حجیم داده‌های زبانی را به زبان‌شناسی پیکره‌ای می‌سپارد (کوهستانی: ۱۳۸۹). زبان‌شناسی پیکره‌ای به‌خودی‌خود چهار محدودیت ذاتی نیز دارد که عبارتند از: ماهیت وقت‌گیر جمع‌آوری و برچسب‌زنی به اقسام کلمه، خطاها و سایر مشخصه‌های زبان. لزوم روزآمد شدن مداوم پیکره‌ها (که در بسیاری از پیکره‌ها مغفول می‌ماند). سوق یافتن خواسته یا ناخواسته زبان‌شناس به نوشتار و نادیده گرفتن گفتار به دلیل اتکاء فراوان به رایانه. نادیده ماندن آن دسته از ویژگی‌های زبان که در پیکره به‌روز نمی‌شوند (این مشکل فلسفی از آنجا ناشی می‌شود که پیکره، نمونه‌ای محدود از جامعه‌ای نامحدود است. سوتیلو: ۲۰۰۵، نقل از کوهستانی: ۱۳۸۹). توبرت (۲۰۰۵) برای مشخص کردن ابعاد زبان‌شناسی پیکره‌ای، ویژگی‌هایی از شیوه مطالعات زبانی در چارچوب زبان‌شناسی پیکره‌ای را بر شمرده است که به شرح زیر است:

تاکید روی معنا: هر جزء از متن (کلمه، عبارت، گروه) حاوی دو وجه صورت و معناست؛ معنا در واقع در قالب صورت، نمایش داده می‌شود. کانون توجه زبان‌شناسی پیکره‌ای، معناست؛ معنا همان چیزی است که از طریق ارتباط گفتاری از فردی به فرد دیگر در جامعه زبانی منتقل می‌شود. معنا در کلام است. اگر بخواهیم بدانیم معنی جزئی از متن چیست باید به سراغ کلام برویم؛ یعنی به اجزای قبلی موجود در متن و ارتباط آن‌ها با جزء مورد مطالعه یا به ارتباط تازه‌ای که جزء مورد نظر با اجزای بعدی متن ایجاد می‌کند، توجه کنیم. زبان‌شناسی پیکره‌ای معنا را در عالم مقال در نظر می‌گیرد. معنا را نمی‌بایست خارج از عالم مقال جستجو کرد. زیرا که هیچ ارتباط مستقیمی بین عالم مقال و جهان خارج وجود ندارد. به عنوان مثال، مصداق ضمائر ارجاعی در عالم مقال مشخص می‌شود. برای زبان‌شناسی پیکره‌ای، معنای متن یا اجزای یک متن، مستقل از قصد و نیت گوینده متن است. چرا که دانستن قصد و نیت گوینده در گونه گفتاری زبان، زمانی که گوینده حاضر است و می‌تواند در رفع ابهام کمک کند، مطرح است. در مثال «سیگار می‌کشید؟» قصد و نیت گوینده می‌تواند عدم آگاهی از مسئله‌ای و سوال کردن درباره آن (اینکه آیا مخاطب سیگار می‌کشد یا نه؟) و یا تعارف کردن باشد. اما آنچه در زبان‌شناسی پیکره‌ای مطرح است معنی اجزای کلام است نه قصد و نیت گوینده.

تاکید روی متن: در زبان‌شناسی پیکره‌ای آنچه بیشتر مطالعه می‌شود، «متن» یعنی صورت نوشتاری زبان است؛ صورتی که قابلیت بازآفریده شدن، تفسیر شدن و شنیده شدن بطور مکرر را دارد و در مقایسه با گفتار که گذرا و ناپایدار است، نوشتار ثابت است.

استفاده از داده‌های واقعی زبان: زبان‌شناسی پیکره‌ای یک رویکرد تجربی در مطالعه ساخت زبان است و با داده‌های واقعی زبان سر و کار دارد. در واقع زبان‌شناسی پیکره‌ای تأکیدی است بر کار کردن روی داده‌های واقعی زبان که از کلام گرفته شده و در یک پیکره جمع‌آوری شده است. این حوزه، ادعای بررسی کلام از طریق تحلیل پیکره زبانی را دارد، و برای این کار از قواعد فراوانی و احتمالات استفاده می‌کند که در حوزه‌های مختلف مانند دستور، تغییرات زبانی یا معنای واژگانی به کار می‌روند. به عنوان مثال، در زبان فارسی می‌توان کاربرد فعل سوم شخص مفرد ماضی را از طریق فراوانی آن در پیکره زبان فارسی بررسی کرد.

در نظر گرفتن بعد در زمانی کلام: هر کلامی حاوی بعد در زمانی است. چیزی که امروز گفته می‌شود عکس‌العملی است برای چیزی که قبلاً گفته شده و همچنین بطور همزمان به چیزی که انتظار داریم در آینده/ فردا گفته شود مربوط است. در زبان‌شناسی پیکره‌ای می‌خواهیم بدانیم چه چیز باعث می‌شود که جزئی از متن تبدیل به رویدادی منحصر به فرد شود، و این زمانی امکان‌پذیر است که نشانه‌های بینا-متنی را پیدا کنیم که این جزء را به اجزای قبلی متن و به متن‌های بعدی مربوط کند؛ زمانی این نشانه‌های بینا-متنی را پیدا می‌کنیم که پیکره دارای ساختار در زمانی باشد.

وجه تمایز زبان‌شناسی پیکره‌ای با سنتی: با آنکه زبان‌شناسی پیکره‌ای از طبقه‌بندی‌های زبان‌شناسی سنتی استفاده می‌کند، اما آن‌ها را مسلم فرض نمی‌کند. خود کلام است که طبقه‌بندی‌های لازم را برای تحقیقات زبانی فراهم می‌آورد و در واقع فقط با پیکره سر و کار دارد. زبان‌شناسی پیکره‌ای یک روش کار نیست، بلکه روش‌های بسیاری را در تجزیه و تحلیل داده‌های پیکره‌ای بکار می‌برد. زبان‌شناسی پیکره‌ای از جهانی‌های زبان (اگر جهانی‌ها را ویژگی‌های هستی‌شناختی در نظر بگیریم و نه مفاهیم نظری) شروع نمی‌کند، بلکه با کاربرد زبان (کنش‌زبانی و نه توانش) سر و کار دارد، زیرا که پیکره زبانی بر اساس کاربرد زبان در امر ارتباط تهیه شده است. زبان‌شناسی پیکره‌ای عناصر درون-مرکز را که به طور صوری بوسیله دستور زبان به هم بسته شده‌اند در نظر گرفته و آن‌ها را اقلام واژگانی یا واحدهای معنایی می‌نامد. آن واحدهای معنایی شامل کلمات، ترکیب‌ها، عبارات و حتی اصطلاحات است. واحدهای واژگانی

همانند کلمات منحصر به فرد در کلام وجود دارند و برای همین است که روش‌های آماری برای کشف آن‌ها در پیکره بکار می‌رود. فراوانی، پارامتری مهم برای کشف الگوهای متناوب زبانی است. ولی به هر حال کار آماری به تنهایی هرگز کافی نیست. اقلام واژگانی علاوه بر ارتباط صوری می‌بایستی به لحاظ معنایی نیز به هم مربوط باشند. اصطلاحات و مدخل‌های واژگانی پیچیده/مرکب، واحدهای معنایی هستند که معنای آن‌ها از طریق تجزیه آن‌ها به واحدهای کوچک‌تری که از آن‌ها تشکیل شده‌اند، قابل دریافت نیست. به عنوان مثال در جمله «سرش را زیر آب کرد» معنای عبارت با دانستن معنای کلمات «سر»، «زیر» و «آب» فهمیده نمی‌شود، بلکه در کل، حاوی معنای خاصی است. مدخل‌های واژگانی، همان‌طور که گفته شده، می‌توانند کلمات یا واحدهای پیچیده معنایی باشند. در واقع مدخل‌های واژگانی تک‌معنا هستند، و همین مسئله است که مفهوم مداخل واژگانی را از مفهوم کلمه (که در بسیاری موارد چند معنا دارد) جدا می‌کند. بنابراین مسئله ابهام معنایی که مشکلات زیادی را برای زبان‌شناسان ایجاد کرده است، برطرف می‌شود. اگر یک واحد واژگانی تک معنا در یک متن تکرار شود، هر وقوع آن مثالی است از نوع واحد واژگانی یکسان. بنابراین هر مثال، نشانه‌ای است از نوعی که بوسیله این واحد واژگانی شناخته می‌شود. بسیاری از انواع واحدهای واژگانی به لحاظ صوری متفاوتند. این تفاوت‌ها اغلب تأثیری در معنا ندارند. ادعای دیگری که زبان‌شناسی پیکره‌ای دارد این است که هر جزء متن از دو منظر مورد بررسی قرار می‌گیرد: ۱- به عنوان مثالی از یک یا چند نوع واحد بالقوه واژگانی ۲- به عنوان رخدادی منحصر به فرد در نظر گرفته شود که معنای آن می‌بایستی از طریق نشانه‌های بینا-متنی که این جزء متن را به متون دیگر کلام مربوط می‌کند، تعیین شود؛ در حالت دوم فراوانی در نظر گرفته نمی‌شود. معنا، دگرگفت است. هر وقت نمادهای واحدهای واژگانی عامل نابسامانی‌های ارتباطی شوند، معنایشان مورد بحث و بررسی قرار گرفته و بعضاً توسط مترادف‌ها جایگزین شده و حتی تعریفی از آن‌ها، هم‌چنانکه در فرهنگ‌های لغت آمده، ارائه می‌شود. آنچه که به عنوان دگرگفت در کلام یافت می‌شود، اجزای متنی است که توسط اعضای جامعه زبانی به عنوان اجزای معنایی در نظر گرفته می‌شود. اما به هر حال یک واحد واژگانی یکسان به طرق مختلف دگرگفت می‌شود. بنابراین زمانی که یک واحد واژگانی دگرگفت می‌شود، می‌توان از دو منظر به آن نگریست: الف - به عنوان رخدادی منحصر به فرد، ب - موردی از نوع واحد واژگانی. از منظر هم‌زمانی، معنای یک نوع واحد واژگانی، تعمیمی است بر همه دگرگفت‌هایی که برای نمودهای این واحد واژگانی یافت می‌شود. از منظر درزمانی، این تاریخ دگرگفت‌های یک جزء متنی تکراری است که به ما می‌گوید منظور از رخداد منحصر به فرد چیست. هیچ معنای حقیقی و ثابتی وجود

ندارد. هرکس می‌تواند یک جزء معنایی را آن‌طور که دوست دارد دگرگفت کند. بنابراین، معنای هر نوع واحد واژگانی موقتی است. اعضای جامعه زبانی به بحث خود در مورد اینکه یک واحد معنایی به چه معنی است ادامه خواهند داد. مسئله واقعیت نیست بلکه مقبولیت است. یک دگرگفت که مقبولیت عام داشته و دوباره استفاده می‌شود، بیشتر مورد بحث است تا دگرگفتی که هیچگاه تکرار نمی‌شود. کلام یک نظام خودارجاعی است. زبان طبیعی تنها نظام رمزگذاری است که در آن نقش اجزای آن بوسیله بحث و جدل‌های درون-کلامی (و نه از خارج از آن) تعیین می‌شود. که این مسئله، زبان طبیعی را از زبان مصنوعی، مانند ریاضیات، جدا می‌کند. وقتی صحبت می‌کنیم به واقعیت خارج کلامی مراجعه نمی‌کنیم، بلکه به آنچه که قبلاً گفته شده ارجاع می‌دهیم. وقتی درباره معنای جزئی از متن بحث می‌کنیم، این کار را در قالب کلام انجام می‌دهیم، نه خارج از آن. بنابراین نظامی دایره‌وار تشکیل می‌شود. (مانند نظام یک فرهنگ لغت که هرگز از دایره ارتباطات معنایی نمی‌توان رهایی یافت). بنابراین هر واحد واژگانی به واحدهای واژگانی دیگر مربوط است. و برای توضیح هر واحد واژگانی که وارد کلام می‌شود باید از واحدهای واژگانی که از قبل در کلام وجود داشته کمک گرفت. کلام حاوی شواهدی است که توسط اعضای جامعه کلامی تهیه شده است. (مثل تجربیات یک شخص نیست.) واقعیت‌های خارج کلامی تنها به عنوان شاهد، وارد کلام می‌شوند. ارتباط بین کلام و واقعیت‌های خارج کلامی، بخشی از نظام زبانی نیست، بلکه این ارتباط باید توسط هریک از اعضای جامعه کلامی برقرار شود. زبان‌شناسی پیکره‌ای به وجود تفاوت بین معنای واژگانی و معنای دایره‌المعارفی قائل نیست. واحدهای واژگانی و اینکه آن‌ها به چه چیز ارجاع می‌دهند، موضوعات کلامی‌اند که توسط اعضای جامعه کلامی ساخته شده‌اند. در مورد موضوعات کلامی مانند «تک شاخ» یا «لیمو» می‌توان گفت: «تک شاخ» به اندازه «لیمو» واقعی است. بستگی به هر یک از اعضای جامعه کلامی دارد که تعیین کنند آیا «لیمو» یا «تک شاخ» جزئی از تجربیات شخصی‌شان است یا نه. زبان‌شناسی همانند علوم طبیعی نیست که محدوده کاری آن، جستجوی حقیقت باشد. به علوم انسانی مربوط است. هیچ معنای حقیقی وجود ندارد. زبان‌شناسی پیکره‌ای عضوی است از جامعه کلامی نه خارج از آن. زبان‌شناسان پیکره‌ای باید یافته‌هایشان را به جامعه کلامی خود عرضه کرده و در مورد مقبولیت آن‌ها بحث کنند. بنابراین عملاً جامعه کلامی، جامعه‌ای دموکراتیک است. هر عنصری حق بحث کردن، کمک کردن و یا رد کردن آنچه را دیگر اعضای جامعه کلامی می‌گویند، دارد. بنابراین کلام خود را سامان می‌بخشد.

۱-۴- کاربرد پیکره در مطالعات زبان‌شناسی و پردازش زبان طبیعی

زبان‌شناسی پیکره‌ای، ابزارهایی بنیادی برای بررسی‌ها، پژوهش‌ها و کاربردهای زبانی متفاوت فراهم می‌آورد. اکنون تنها به برخی از این موارد اشاره می‌کنیم (عاصی: ۱۳۸۲): یکی از مهم‌ترین کاربردهای پیکره در پردازش زبان طبیعی^۱ است. مهم‌ترین دستاورد این حوزه، درک و بازشناسی گفتار بوده است که تنها با بهره‌گیری از پیکره‌های بزرگ امکان پذیر شده است. از دیگر کاربردهای پیکره، استفاده از پیکره-ها در فرهنگ‌نگاری است؛ اکنون هیچ پروژه فرهنگ‌نگاری پیشرفته‌ای نمی‌توان یافت که از پیکره‌های زبانی و پایگاه‌های داده‌های زبانی بهره‌گیری نکند. پیکره‌ها ابزاری کارا و ناگزیر در کار تدوین فرهنگ‌ها به شمار می‌روند. (هارتمن: ۱۹۹۴، نقل از عاصی: ۱۳۸۲). نمونه چنین کاربردی در زبان فارسی «واژگان گزیده زبان‌شناسی» است (عاصی: ۱۹۹۴، عاصی: ۱۳۷۳ و عاصی و عبدعلی: ۱۳۷۵، نقل از عاصی: ۱۳۸۲). ایجاد پایگاه‌های داده‌های زبانی نیز جنبه‌ای دیگر از کاربرد پیکره‌های زبانی است که نمونه‌های متعدد آن را هم‌اکنون در سراسر جهان به صورت پیوسته یا ناپیوسته در اختیار داریم. طرح‌های بررسی واژه‌های همایند^۲ در زبان‌های گوناگون با کمک پیکره‌های زبانی اجرا شده است. نمونه مهم و موفق آن «فرهنگ واژه‌های همایند BBI برای زبان انگلیسی است. برنامه‌های پایشگری زبان برای پیگیری و ردگیری تحولات زبانی نیز از امکانات پیکره‌های زبانی سود می‌برند. این گونه پیکره‌ها را پیکره پویا یا پیکره پایشگر می‌نامند (کندی: ۱۹۹۸، نقل از عاصی: ۱۳۸۲). همه طرح‌های ترجمه ماشینی، به گونه‌ای از پیکره‌های زبانی سود می‌برند. استفاده از پیکره‌های زبانی در مطالعات ترجمه به عنوان ابزاری برای مترجمان، یکی از حیطه‌های سریع‌اً رو به رشد و جالب در کار ترجمه تجربی شده است. پیکره زبانی ما را قادر می‌سازد تا توجه خود را بر روی زبان به صورتی که در واقع در ترجمه به کار می‌رود معطوف کنیم و در نتیجه ما را قادر می‌سازد تا معین کنیم در ترجمه، به عنوان یک نوع از متن، چه چیزی محتمل و طبیعی است. به این صورت می‌توانیم ترجمه‌ها را با صورت‌های دیگر متن‌ها مقایسه کنیم. روش‌های مبتنی بر پیکره زبانی همچنین این توانایی را به ما می‌دهد تا بر روی ترکیب‌های واژگانی، نحوی، و ویژگی‌های کلامی در هنگام مقایسه شمار زیادی از ترجمه‌ها به زبان‌های مختلف که توسط مترجمان گوناگون و در محیط‌های اجتماعی-فرهنگی متفاوت، و در زمان‌های مختلف صورت گرفته، متمرکز شویم. از دیگر کاربردهای زبان‌شناسی پیکره‌ای، استفاده از پیکره‌ها در تدریس و یادگیری زبان، تشخیص گویش‌ها و

¹ natural language processing

² collocations

تنوعات زبانی، معنی‌شناسی، تحلیل کلام، زبان‌شناسی اجتماعی، زبان‌شناسی حقوقی، تجزیه و تحلیل ژانرهای ادبی و مطالعات دستوری است. با توجه به کارایی‌های فزاینده پیکره‌های زبانی و گسترش رشته زبان‌شناسی پیکره‌ای، می‌توان گفت که تنها در سه دهه گذشته، این رشته رشد اعجاب‌انگیزی داشته است و به روش‌های نوین، اندیشه‌ها و کارکردهای تازه با کمک دستگاه‌های الکترونیک و رایانه دست یافته است. با همه این‌ها، هنوز بزرگ‌ترین و مهم‌ترین دستگاہی که می‌تواند از میان انبوه واقعیت‌ها قوانین را استخراج و تدوین کند، همان ذهن انسان است (اسوارت ویکل: ۱۹۹۲، نقل از عاصی: ۱۳۸۲). مهم‌ترین مزیت استفاده از پیکره‌های زبانی این است که در مطالعات زبانی، محقق مجبور نیست به شم زبانی خود یا دیگر گویشوران زبان تکیه کند، بلکه می‌تواند از حجم عظیمی از داده طبیعی و اصیل که حاصل گفتار یا نوشتار گویشوران است، جهت بررسی فرضیه زبانی خود استفاده کند و فرضیه خود را با داده‌های تجربی محک بزند.

۱-۵- معرفی برخی از پیکره‌های زبانی فارسی

۱-۵-۱- پیکره متنی زبان فارسی: این پیکره، در واقع توسعه‌یافته «پیکره بی‌جن‌خان» است. پیکره بی‌جن‌خان، مجموعه‌ای است از متون فارسی شامل بیش از ۲ میلیون و ۶۰۰ هزار کلمه که با ۵۵۰ نوع برجسب اجزای واژگانی کلام^۱، برجسب‌گذاری شده‌اند. این پیکره که در پژوهشکده پردازش هوشمند علائم تهیه شده است، همچنین شامل بیش از ۴۳۰۰ برجسب موضوعی چون سیاسی، تاریخی و ... برای متون است (بی‌جن‌خان: ۱۳۸۳). پیکره متنی استاندارد زبان فارسی مجموعه‌ای از متون نوشتاری و گفتاری زبان فارسی به صورت رسمی است که از منابع واقعی همچون روزنامه‌ها، سایت‌ها و مستندات از قبل تایپ شده، جمع‌آوری شده، تصحیح گردیده و برجسب خورده است. حجم این دادگان حدوداً ۱۰۰ میلیون کلمه است و از منابع مختلف تهیه گردیده و دارای تنوعات بسیار زیادی است. پیکره متنی زبان فارسی دارای قابلیت‌ها و ویژگی‌هایی است که شامل: جمع‌آوری و سازمان‌دهی متون نوشتاری و گفتاری رسمی زبان فارسی با حجم ۱۰۰ میلیون کلمه. ویرایش نیمه‌خودکار اولیه متون. برجسب‌دهی نحوی-معنایی کلمات برای ۱۰ میلیون کلمه با استفاده از ۸۸۲ برجسب به صورت دستی توسط دانشجویان رشته زبان‌شناسی بر اساس دستورالعمل تهیه نویسه-

¹ Part Of Speech (POS)

های Unicode و XML برای پرونده‌های متنی دادگان. امکان برچسب‌دهی گروه‌های نحوی. طبقه‌بندی هر پرونده بر حسب موضوع و منبع آن. پوشش موضوعات مختلف سیاسی، اجتماعی، اقتصادی، فرهنگی، مجهز به یک نرم افزار آماری برای محاسبه و استخراج ویژگی‌های زبانی از قبیل: توزیع احتمالی مشروط، واژگان بسامدی، شناسایی هم-نگاره‌ها، هماینها، مطابقتها و ترتیب قاموسی با امکان گزارش‌گیری. طراحی یک زبان جستجوی هوشمند. اولین و تنها پیکره متنی حجیم در زبان فارسی، است (بی‌جن‌خان و همکاران: ۲۰۱۱). برخی از کاربردهای پیکره متنی فارسی شامل: استفاده در تعلیم مدل زبانی^۱ برای سیستم‌های بازشناسی گفتار پیوسته فارسی با دایره لغات بزرگ (LVCSR) و استفاده در پروژه‌های مربوط به پردازش زبان طبیعی است.

۱-۵-۲-

پایگاه دادگان زبان فارسی: این پیکره، مجموعه‌ای نرم‌افزاری برای ذخیره، پردازش و ارائه داده‌های زبانی فارسی است. این پایگاه دربرگیرنده پیکره‌های گوناگونی از زبان فارسی است که با وجود حجمی عظیم و با گستردگی و گوناگونی‌های بسیار، دارای ساختاری بسامان و منطقی است و امکان هرگونه جستجو و دستیابی سریع به آگاهی‌های مورد نیاز را در هر زمان فراهم آورده است. پیکره‌های این پایگاه می‌توانند همواره روزآیند شوند. پایگاه دادگان زبان فارسی فراگیر و متنوع است. در واقع فراتر از یک یا چند پیکره خاص است و کاربران بر پایه نیاز و هدف پژوهشی خود می‌توانند پیکره مناسب را از آن برگزینند. حتی پژوهندگان می‌توانند پیکره‌های اختصاصی خود را وارد پایگاه کنند و تحلیل‌ها و فهرست‌گیری‌های مورد نظر خود را انجام دهند. پایگاه دادگان زبان فارسی تنها مجموعه‌ای از مواد خام زبانی نیست بلکه دارای متن‌های نشانه‌گذاری شده از جمله شناسنامه متن، برچسب‌های دستوری، آوایی، ریشه‌ای و معنایی است که همواره افزایش می‌یابد. این دادگان مجهز به نرم‌افزارهای اختصاصی جستجو، تقطیع و تحلیل متن است که می‌تواند انواع فهرست‌های واژگانی، بسامدی و آماری را ارائه کند. منابع مورد استفاده در این پایگاه از گونه‌های نوشتاری با استفاده از متن‌های معتبر و با رعایت معیارهای مختلف نمونه‌گیری شده است و البته هیچ‌گونه محدودیت و امساک در مورد آثار مهم ادبی و نویسندگان سرشناس و بویژه صاحب سبک و تاثیرگذار اعمال نمی‌شود. فهرست‌های مفصلی از همه منابع مهم نظم

¹ language model

و نثر فارسی معاصر فراهم شده است. این فهرست‌ها به طور جداگانه برای آثار شعری، داستانی، غیرداستانی، نمایشنامه و فیلمنامه، ادبیات کودکان، نشریه‌های ادورای و مجلات علمی، تخصصی و ادبی فراهم گردیده است. شمار آثاری که در این فهرست‌ها قرار گرفته‌اند، بیش از یک هزار و پانصد عنوان شده است که پس از بررسی و کنار گذاشتن موارد مشابه، بیش از پانصد عنوان برای درونداد پایگاه داده‌ها برگزیده شده است. حدود ۴۵۰ اثر داستانی و غیر داستانی نثر، ۲۵۰ اثر شعری از شاعران معاصر، بیش از ۸۰ عنوان مجله و نشریه علمی، ادبی و تخصصی، نزدیک به ۳۰۰ عنوان نمایش‌نامه و فیلم‌نامه، و ۲۰۰ عنوان ادبیات کودک، چندین عنوان روزنامه و نشریه خبری، برخی از کتاب‌های درسی دانشگاهی و دبیرستانی، برخی از کتاب‌های دبستانی، نامه‌های اداری و بخشنامه‌ها، مجموعه کامل قوانین و مقررات، نشریه‌ها و جزوه‌های پراکنده، پوسترها، دیوارنوشته‌ها و مانند این‌ها از جمله این متون هستند (عاصی: ۱۳۸۴).

۱-۵-۳-

پیکره وابستگی نحوی زبان فارسی: یکی از نیازهای اساسی در پردازش رایانه‌ای زبان و فهم خودکار محتوا و معنی، تحلیل نحوی جملات است. چنین تحلیلی علاوه بر آنکه به خودی خود، یاری‌گر رایانه در شناخت مفهوم است، پیش‌پردازشی برای تحلیل معنایی زبان نیز تلقی می‌شود. پیکره وابستگی نحوی زبان فارسی، مجموعه‌ای است شامل ۳۰۰۰۰ جمله برچسب‌خورده با اطلاعات نحوی و ساخت‌واژی. این پیکره که نخستین پیکره وابستگی زبان فارسی است، می‌تواند به عنوان زیرساختی اساسی در پردازش رایانه‌ای زبان فارسی به کار رود. همچنین اطلاعات موجود در این پیکره می‌تواند در پژوهش‌های زبان‌شناختی و آموزش و یادگیری زبان فارسی مورد استفاده قرار گیرد. مهم‌ترین دلایل استفاده از دستور وابستگی در این پیکره نحوی عبارتند از: نتایج رضایت‌بخش در یادگیری خودکار و سازگاری مناسب با طبیعت زبان‌های بی‌ترتیب، همچون زبان فارسی (منظور از بی‌ترتیب بودن زبان فارسی این است که برای اکثر جملات، ترتیب خاصی برای جملات نمی‌توان در نظر گرفت. مانند جمله «من در مدرسه کتاب را به علی دادم» که به صورت‌های دیگر نیز می‌توان این مفهوم را بیان کرد: «من در مدرسه به علی کتاب را دادم»، «من به علی در مدرسه کتاب را دادم» و «من کتاب را به علی در مدرسه دادم». ویژگی‌های پیکره وابستگی نحوی زبان فارسی شامل: جملات پیکره برگرفته از منابع مختلفی از متون فارسی معاصر هستند. تمامی جملات دارای

برچسب روابط نحوی (بر مبنای دستور وابستگی) از قبیل فاعل، مفعول، مسند، مضاف‌الیه، بدل و.... هستند. تمامی جملات دارای برچسب اطلاعات ساخت‌واژی (برچسب اجزای واژگانی کلام (POS)) از قبیل فعل، اسم، صفت، قید، ضمیر و.... هستند. جملات توسط تیمی از زبان‌شناسان مجرب برچسب خورده‌اند و در چند مرحله بازبینی شده‌اند. داده‌های پیکره به صورت تصادفی، به داده‌های یادگیری (۸۰٪)، آزمون (۱۰٪) و ارزیابی (۱۰٪) تقسیم شده است. تعداد کل جملات ۲۹۹۸۲، تعداد کل واژه‌ها ۴۹۸۰۸۱، تعداد واژه‌های منحصر به فرد ۳۷۶۱۸، میانگین طول هر جمله ۶۱/۱۶، تعداد افعال منحصر به فرد ۴۷۸۲ و میانگین حضور هر فعل ۶۷/۱۲ است (رسولی و همکاران: ۲۰۱۳).

۱-۵-۴-

پیکره همشهری: مجموعه همشهری پیکره‌ای است حاوی ۳۱۸ هزار سند مربوط به اخبار سال‌های ۱۳۷۵ تا ۱۳۸۶ که با خزش (Crawl) وب‌سایت همشهری و چندین مرحله پیش‌پردازش و برچسب‌گذاری حاصل آمده است. همه اسناد مجموعه همشهری دارای برچسب «Cat» هستند که نشان می‌دهد هر سند در چه رده‌ای است (اقتصادی، سیاسی و...). (آل احمد و همکاران: ۲۰۰۹).

۱-۵-۵-

پیکره فارس‌نت: فارس‌نت (وردنت عمومی زبان فارسی) پایگاه دانشی است که حاوی اطلاعات درمورد واژه‌ها و ترکیبات زبان (مفاهیم)، اطلاعات نحوی آن‌ها و روابط معنایی میان آن‌ها است. نسخه اول فارس‌نت شامل بیش از ۱۷ هزار مدخل واژگانی از مقوله‌های اسم، فعل و صفت است. روابط تحت پوشش در این نسخه روابط درون‌مقوله‌ای مطرح در وردنت انگلیسی (نسخه ۲ و ۱) می‌باشد و قابلیت اتصال به وردنت‌های دیگر از طریق نگاشت به وردنت پرینستون نسخه ۳۰۰ را نیز داراست. نسخه دوم فارس‌نت شامل بیش از ۳۰ هزار مدخل واژگانی از مقوله‌های اسم، فعل، صفت و قید است. علاوه بر روابط درون-مقوله‌ای مطرح در وردنت انگلیسی (نسخه ۲ و ۱)، پنج رابطه میان-مقوله‌ای نیز مفاهیم را بهم پیوند می‌دهد و علاوه بر ویژگی‌های در نظر گرفته شده برای واژه‌ها، ویژگی‌های نحوی، ساخت‌واژی و آوایی به واژه‌ها و قاب و ساختار آرگومانی به افعال افزوده شده است. این وردنت نیز قابلیت اتصال به وردنت‌های دیگر را از طریق نگاشت به وردنت پرینستون نسخه ۳۰۰ داراست. مجموعه فارس‌نت توسط آزمایشگاه پردازش زبان طبیعی دانشگاه شهید بهشتی

و با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) تهیه شده است (شمس‌فرد و همکاران: ۲۰۱۰).

۱-۵-۶-

پیکره نقش‌های معنایی زبان فارسی: این پیکره که حدود ۳۰,۰۰۰ جمله از زبان فارسی معاصر را شامل می‌شود، به صورت دستی برچسب‌گذاری شده است. این پیکره بر اساس مفهوم نقش‌های معنایی فیلمور، لایه‌ای از اطلاعات مربوط به رابطه محمول موضوع را به ساخت نحوی پیکره وابستگی اضافه می‌کند. در این مجموعه، افعال، اسم‌ها و صفت‌های گزاره‌ای به عنوان محمول‌های جمله در نظر گرفته شده و بنا بر نوع رویدادشان، در جمله تعیین ظرفیت شده‌اند. خروجی این پیکره بر اساس الگوی همایش زبان‌شناسی رایانه‌ای و پردازش زبان طبیعی (CoNLL) آماده شده است. نقش‌های معنایی مورد استفاده در برچسب-زنی معنایی شامل دو گروه برچسب‌های موضوعی و برچسب‌های نقشی هستند. برچسب‌های موضوعی، ساخت ظرفیتی محمول‌های جمله را به دست می‌دهند و برچسب‌های نقشی، افزوده‌های توصیف‌گر فعل یا کل جمله را شامل می‌شوند. تعداد نقش‌های معنایی ۲۷ مورد و تعداد موضوع‌های نقشی ۱۵ مورد است. دو برچسب وجه‌نما و نفی هم به عنوان برچسب نقشی مورد استفاده قرار گرفته است (میرزایی و مولودی: ۱۳۹۳).

۱-۵-۷-

پیکره تشخیص خودکار جنسیت: این پیکره شامل دو بخش اصلی است که عبارتند از: ۱- بخش متون رسمی که با مشخص کردن جنسیت نویسندگان متون داستانی موجود در پیکره بی‌جن‌خان و داستان‌های دیگر برگرفته از اینترنت به دست آمده است. ۲- بخش متون غیررسمی. از رویکرد وب برای پیکره استفاده شده است. از نظرات کاربران در سایت «هلو کیش» نیز استفاده شده است. برای استخراج نظرات مرتبط با نظردهندگان زن و مرد، فهرستی از اسامی فارسی زن و مرد استفاده شده است که نام نویسنده مدنظر با این فهرست تطبیق داده شده است و نظرات برحسب این فهرست تفکیک شده و در دو دسته زن و مرد قرار گرفته است (مرادی و بحرانی: ۱۳۹۴).

۱-۵-۸-

پیکره واژگان فارسی دارای برچسب قطبیت: این پیکره که در آزمایشگاه سیستم‌های هوشمند اطلاعات دانشگاه تهران تهیه شده است، از دو مجموعه داده تشکیل شده است: ۱- مجموعه استخراج شده از صفت‌های برچسب‌خورده فارسی؛ این مجموعه از روی مجموعه صفت‌های زبان فارسی استخراج شده از فارس‌نت ساخته شده است. هر ورودی در این

مجموعه می‌تواند برچسب مثبت، منفی و یا خنثی داشته باشد. برای این کار بیش از ۳۵۸۸ صفت استخراج و توسط چهار ارزیاب مستقل ارزیابی شده است. پس از تجمیع نظرات چهار ارزیاب، سه مجموعه مثبت، منفی و خنثی استخراج شده است که به ترتیب شامل ۹۶۸ ۹۶۲ ۱۵۷۲ کلمه می‌باشند. از بین صفتهای ارزیابی شده، تنها صفتهایی در مجموعه باقی مانده‌اند که یا توسط هر چهار ارزیاب برچسب یکسان گرفته‌اند و یا سه ارزیاب در مورد صفت نظر یکسان داشته‌اند و ارزیاب چهارم به این صفت برچسب مخالف نداده است. ۲- مجموعه صفت، فعل و اسم: این مجموعه از روی مجموعه صفتهای فعل‌ها و اسم‌های موجود در فارسی‌نت استخراج شده‌اند. به هر کلمه از این مجموعه توسط یک روش مبتنی بر یادگیری ماشینی نیمه-ناظر یک مقدار عددی نسبت داده می‌شود. این عدد در واقع تعیین‌کننده میزان قطبیت هر کلمه است. به کلمات مثبت، عددی بزرگتر از صفر و به اعداد منفی، عددی کوچکتر از صفر نسبت داده می‌شود. در این مجموعه، کلمات خنثی به صراحت تعیین نمی‌شوند و می‌توان کلمات خنثی را بر اساس یک حد آستانه بین کلمات مثبت و منفی تعیین کرد. این مجموعه شامل ۳۵۸۸ صفت ۴۰۷۳ فعل و ۷۳۲۵ اسم می‌باشد. قابل ذکر است که کلیه کلمات از روی نسخه ۱ فارسی‌نت استخراج شده‌اند (دهداربهبهانی و همکاران: ۲۰۱۴).

۱-۵-۹- پیکره چندزبانه رایانامه‌ها: این پیکره برای تشخیص ریسمان‌های گفتگوی چندزبانه در آزمایشگاه سیستم‌های هوشمند اطلاعات دانشگاه تهران تهیه شده است. نام این پیکره «Multilingual-BC3» است که در حقیقت یک پیکره ساختگی چندزبانه است که حاصل ترجمه بخشی از پیکره تک‌زبان BC3، توسط عامل انسانی است. پیکره اولیه BC3 به صورت تک‌زبان و در زبان انگلیسی توسط آزمایشگاه هوش محاسباتی در دانشگاه British Columbia ساخته شده است. این پیکره، یک زیرمجموعه از پیکره W3C است که دارای برچسب‌های معنایی، نظیر حالت گفتار در سطح جمله و برچسب خلاصه‌سازی گفتگوها است. پیکره ConThread-BC3 یک نسخه از BC3 است که در آن، برچسب‌های نشان‌دهنده ساختار ریسمان‌های گفتگو و همچنین اطلاعات مربوط به برچسب متن اصلی و متن نقل‌قول رایانامه‌ها اضافه شده است. پیکره Multilingual-BC3، در دو نسخه تهیه شده که نسخه اول، گونه‌ای چندزبان از نسخه اولیه BC3 و نسخه دوم گونه‌ای چندزبان از

ConThread-BC3 است. شایان ذکر است که اطلاعات برچسب‌های موجود در نسخه‌های تک‌زبانه مستقل از زبان بوده و قابل گسترش به Multilingual-BC3 خواهند بود (دهقانی و همکاران: ۲۰۱۳).

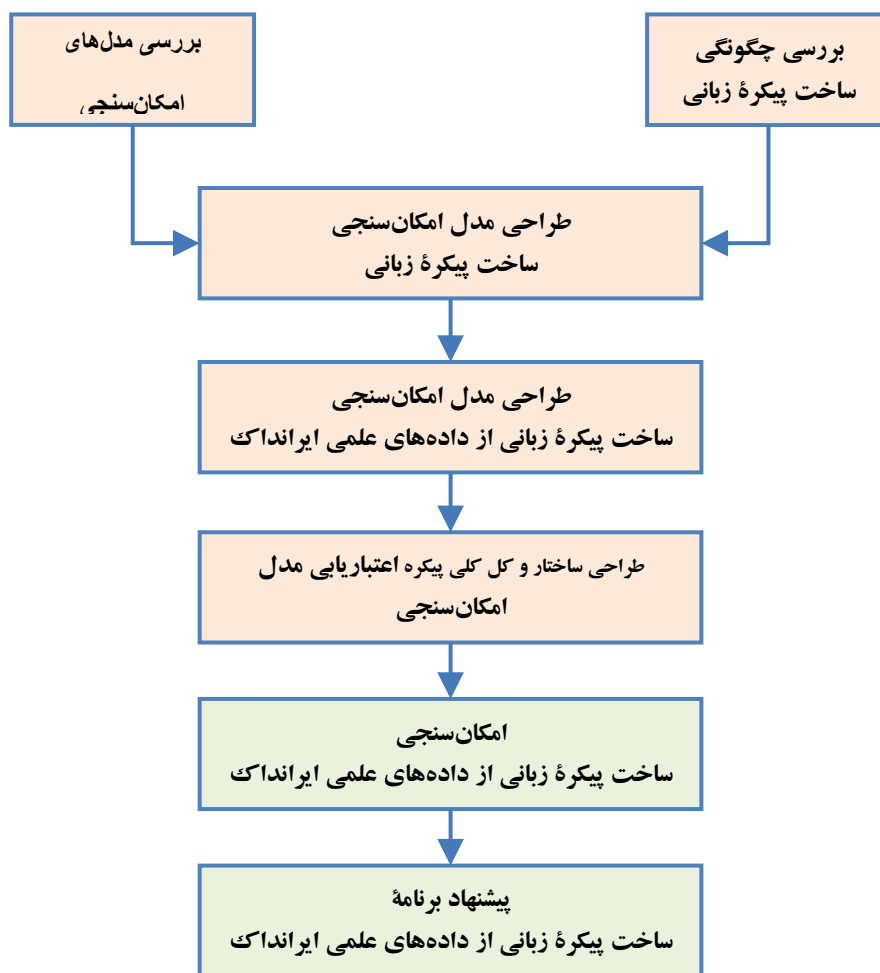
۶-۱- ضرورت انجام پژوهش

تهیه یک پیکره زبانی مبتنی بر داده‌های علمی ایرانداک از سه نظر حائز اهمیت است: ۱- ایرانداک را در شمار پژوهشگاه‌ها و دانشگاه‌هایی قرار می‌دهد که پیکره زبانی مفید تهیه کرده‌اند و پژوهشگران می‌توانند از این پیکره خاص، که پیکره‌ای مبتنی بر داده‌های علمی ایرانداک است، استفاده کنند، ۲- با توجه به حجم عظیم داده‌های علمی در ایرانداک (نوعاً پایان‌نامه‌ها)، با توجه به اینکه این اطلاعات علمی عظیم تنها در اختیار ایرانداک است، ساختن چنین پیکره‌ای کاملاً منحصر به فرد و بی‌رقیب است؛ تاکنون پیکره‌ای علمی مبتنی بر چنین داده‌های عظیمی ساخته نشده است و می‌تواند مورد استفاده جامعه دانشگاهی قرار گیرد و ۳- می‌توان بسیاری از پژوهش‌های حوزه پردازش متن (مانند انواع طبقه‌بندی داده‌ها، استخراج اطلاعات گوناگون، ریشه‌یابی کلمات، تقطیع کلمات، پردازش نحوی جملات و...) را با استفاده از این پیکره انجام داد. برای تهیه چنین پیکره‌ای ابتدا لازم است از تجربیات محققین دیگر که به ساختن پیکره پرداخته‌اند، استفاده کرد و به بررسی مراحل، ابزارها، نیروهای تخصصی لازم، هزینه و... که برای تهیه چنین پیکره‌ای لازم است، پرداخت. بنابراین، در این پژوهش به امکان‌سنجی تهیه پیکره زبانی مبتنی بر داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، پرداخته می‌شود. ضرورت انجام امکان‌سنجی تهیه پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، جمع‌آوری اطلاعات در زمینه مراحل، ابزارهای مورد نیاز، تیم‌های تخصصی لازم، برآورد هزینه و... جهت ساخت پیکره‌ای مطلوب و بسامان مبتنی بر داده‌های ایرانداک است تا بتوان درباره ساخت چنین پیکره‌ای در ایرانداک بهتر تصمیم‌گیری کرد و در صورت داشتن امکانات مالی و نیروی متخصص و ابزارهای لازم، در طرح/ طرح‌های پژوهشی دیگر، اقدام به ساخت چنین پیکره‌ای کرد. بنابراین، در این طرح پژوهشی، پیکره‌ای تهیه نخواهد شد، بلکه امکان‌سنجی تهیه پیکره انجام می‌شود و در طرح‌های پژوهشی آینده، در صورت داشتن امکانات کافی، پیکره تهیه خواهد شد تا بتوان پژوهش‌های مربوط به پردازش متن را در آینده به کمک پیکره‌ای که تهیه خواهد شد، انجام داد.

۱-۷- سوال پژوهش

در این پژوهش به این مسئله پرداخته می‌شود که برای تهیه یک پیکره زبانی بر اساس داده‌های علمی ایرانداک، چه مراحل لازم است طی شود؟ چه ابزارها و تیم‌های تخصصی لازم است؟ هزینه‌های مربوط به مراحل و ابزارها و تیم‌های تخصصی به چه صورت می‌تواند باشد؟ مولفه‌ها و ابعاد ساخت پیکره زبانی مبتنی بر داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) چیست؟ مولفه‌ها و ابعاد مدل امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) چه چیزهایی می‌تواند باشد؟

برای امکان‌سنجی تهیه پیکره زبانی از داده‌های ایرانداک باید مراحل زیر طی شود و با طی مراحل زیر می‌توان درباره تهیه پیکره در ایرانداک بهتر تصمیم‌گیری کرد.



در این طرح پژوهشی ابتدا در فصل دوم، به پیشینه پژوهش، که در واقع مرور مطالعات انجام شده در حوزه ساخت پیکره زبانی است، پرداخته می‌شود. فصل سوم، مبانی نظری ساخت پیکره زبانی است. در فصل چهارم، به امکان‌سنجی ساخت پیکره زبانی پرداخته می‌شود. فصل پنجم، امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی ایرانداک است و در ادامه پیشنهاد برنامه ساخت پیکره زبانی از داده‌های علمی ایرانداک ارائه می‌شود و در نهایت، فصل ششم، نتیجه‌گیری است.

فصل دوم

پیشینه پژوهش

۲-۱- مقدمه

در این فصل مطالعات انجام شده در حوزه ساخت پیکره زبانی، نوعاً بیشتر پیکره‌های متنی، مورد بررسی قرار خواهند گرفت، زیرا هدف از این امکان‌سنجی، بررسی امکان ساخت پیکره‌ای متنی بر اساس داده‌های علمی مکتوب در ایرانداک است.

۲-۲- مرور پژوهش‌های پیشین ساخت پیکره زبانی

واینه (۲۰۰۵) آراء نویسندگان و محققین سرشناس حوزه ساخت پیکره را ارائه می‌کند؛ آن‌ها پیشنهادات کاربردی برای ساخت پیکره ارائه می‌دهند. راهنمایی‌های این محققین می‌تواند مورد استفاده کسانی قرار گیرد که در مرحله ساخت پیکره زبانی هستند. از موضوعاتی که در این کتاب مورد بررسی قرار می‌گیرد تعاریف کلی مربوط به پیکره، اضافه کردن حاشیه‌نویسی زبانی، نوع داده‌های مورد استفاده و پیکره آوایی را می‌توان ذکر کرد. این کتاب، کتابی کاربردی است که به زبان ساده به بررسی مراحل ساخت پیکره زبانی می‌پردازد. عاصی (۱۳۸۲) نیز پس از اشاره به تاریخچه شکل‌گیری رشته زبان‌شناسی پیکره‌ای، به رویکردهای نظری و کاربردی زبان‌شناسی پیکره‌ای، انواع پیکره‌های زبانی، نشانه‌گذاری پیکره‌ها، سازمان‌دهی آن‌ها، پایگاه‌های داده‌های زبانی و نقش رایانه در آن‌ها پرداخته است. سپس، برخی از پیکره‌های مهم زبان انگلیسی و فارسی و نیز چند پروژه فرهنگی‌نگاری را که با بهره‌گیری از پیکره‌های زبانی انجام گرفته‌اند، معرفی می‌کند.

پیکره‌ها می‌توانند تخصصی باشند و برای اهداف خاصی ساخته شده باشند، در این راستا می‌توان به پیکره بالینی روبرتز و همکاران (۲۰۰۹) اشاره کرد. آن‌ها، ساخت یک پیکره بالینی که نوع حاشیه‌نویسی آن، معنایی است را توضیح می‌دهند. پیکره مذکور حاوی متون حوزه بالینی است که برای توسعه و ارزیابی سیستم‌های خودکار بازیابی اطلاعات بالینی مهم مربوط به بیماران، استفاده می‌شود. در این مقاله، جزئیات نمونه‌گیری از متون توضیح داده شده است؛ نمونه‌گیری از ۲۰۰۰۰ سند مربوط به بیماران سرطانی انجام شده است. همچنین روش حاشیه‌نویسی، نحوه توزیع حاشیه‌نویسی‌ها در پیکره و استفاده از پیکره برای توسعه سیستم بازیابی اطلاعات نیز توضیح داده شده است. حاصل پژوهش، پیکره‌ای است غنی که حاوی متون همراه با حاشیه‌نویسی معنایی جهت پردازش متون بالینی است. آن‌ها همچنین معتقدند، از آنجایی که ساخت پیکره مذکور با جزئیات کامل توضیح داده شده است، قابل استفاده کسانی است که قصد دارند پیکره‌هایی متنی با کیفیت بالا همراه با حاشیه‌نویسی معنایی در حوزه پزشکی، بسازند.

همانگونه که پیش از این، در قسمت کاربرد پیکره، توضیح داده شد، یکی از کاربردهای پیکره در امر آموزش است. کلود توریدا (۲۰۱۶) نیز به توضیح مرحله به مرحله ساخت پیکره تخصصی^۱ و تهیه فهرست واژگان حاشیه‌نویسی شده و مبتنی بر فراوانی واژگان در پیکره می‌پردازد. مثال‌ها از پروژه «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته است، استخراج شده است. این مراحل شامل انتخاب متون آموزشی، حذف واژگانی که قاموسی^۲ نیستند (مانند حروف اضافه، حروف ربط و....، البته طی این مرحله جداول، شکل‌ها و... نیز با بررسی حذف می‌شوند)، تجزیه و تحلیل متن با استفاده از نرم‌افزار AntConc، ایجاد فهرست فراوانی واژه‌ها، توسعه فهرست واژگان حاشیه‌نویسی شده که خود شامل: تعیین مقوله نحوی (POS) واژگان موجود در فهرست، اضافه کردن تعریف واژگان، باهم آبی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به کار رفته است، می‌باشد. وی در نهایت توضیح مختصری درباره چگونگی استفاده از فهرست واژگان مذکور در کلاس‌های درس، ارائه می‌دهد. السولیتی و اتول (۲۰۰۶) معتقدند پیکره‌ها منبع مهم برای تدریس و تحقیق به شمار می‌روند. عربی فاقد منابع کافی در این زمینه است، بنابراین در این پژوهش روند تهیه پیکره‌ای برای عربی که در آن وضعیت عربی معاصر و نیازهای کاربران پیکره لحاظ شده باشد، مورد مطالعه قرار می‌گیرد. این مقاله در واقع گزارشی است از نتایج یک تحقیق ساخت پیکره، براساس نیاز معلمانی که زبان عربی را به عنوان زبان خارجی تدریس می‌کنند و همچنین نیاز مهندسين زبان نیز لحاظ شده است. بررسی‌های انجام شده نشان می‌دهد که باید انواع گوناگون متن در حجم وسیع در پیکره گنجانده شود، زیرا پیکره‌های موجود از نظر نوع منبع و ژانر بسیار محدود هستند و باید پیکره‌ای مبتنی بر عربی معاصر که به سهولت و به طور رایگان در دسترس باشد، تهیه شود. السولیتی و اتول، نمونه اولیه‌ای از این پیکره تحت عنوان «the corpus of contemporary Arabic (CCA)» تهیه کرده‌اند و موارد مذکور را در این پیکره لحاظ کرده‌اند.

پیکره‌ها را می‌توان از منابع گوناگون مانند کتاب‌ها، مجلات، صفحات وب و... تهیه کرد. کامیابی گل و همکاران (۱۳۹۷) به معرفی و توصیف چگونگی ساخت پیکره مقالات علمی-پژوهشی که نگارندگان بوجود آورده‌اند، می‌پردازند. برای ایجاد پیکره، ابتدا نرم‌افزار پیکره‌ساز طراحی و ساخته شده است. این نرم‌افزار انواع فرمت‌های متنی را پشتیبانی می‌کند، همچنین می‌توان متغیرهای ساخت پیکره را از قبل تعیین نمود، به عنوان مثال می‌توان مشخص کرد که حداقل تعداد token فایل برای حضور یک متن در پیکره چه عددی باشد. سپس مجموعه مقالات علمی-پژوهشی اعضای هیات علمی دانشگاه فردوسی مشهد

^۱ specialized corpus

^۲ Content words

جمع‌آوری شده است. مجموع مقالات مشتمل بر ۷,۱۵۴,۲۰۲ کلمه و ۱۱۰۰ مقاله است. کل پیکره در فایل‌های مجزا به جملات تشکیل‌دهنده‌اش تجزیه شده است، ریشه کلمات استخراج گردیده و سپس برچسب نحوی کلمات زده شده است. علاوه بر امکان استخراج اطلاعات به‌طور مستقیم، نرم‌افزار جانبی دیگری با کاربرد ساده‌تر برای استخراج اطلاعات آماری نیز طراحی و به آن افزوده شده است تا کاربران غیرتخصصی هم بتوانند از آن استفاده کنند. نگارندگان معتقدند، بر اساس نتایج بدست آمده، پیکره ساخته شده قابلیت بسیار بالایی برای داده‌کاوی و استفاده در تمام پژوهش‌هایی که بر روی متون علمی انجام می‌شود را دارا است. با استفاده از پیکره می‌توان توصیفی داده-محور از نحوه کاربرد زبان توسط گروه‌های مختلف کاربران زبانی ارائه کرد.

پیکره‌ها انواع گوناگونی دارند، پیکره‌ها می‌توانند آوایی باشند و برای اهداف خاصی تهیه شده باشند. در فصلی از کتاب «آواشناسی پیکره‌ای» نیز، دوران و همکاران (۲۰۱۴) به فرآیند ساخت پیکره‌های آوایی می‌پردازند. در این راستا آن‌ها پس از تعریف پیکره آوایی، به بحث پیرامون اجزای مهم ساخت چنین پیکره‌هایی می‌پردازند؛ این اجزا شامل: ذخیره پیکره، چگونگی به اشتراک گذاشتن و استفاده مجدد از پیکره، سوال‌های مربوط به نمایندگی و اندازه پیکره، انتخاب داده خام و حاشیه‌نویسی پیکره است. آن‌ها در نهایت، نظریه‌های مربوط به فرآیند ساخت پیکره را مورد بحث قرار می‌دهند.

تهیه پیکره متنی برای زبان فارسی با مشکلاتی همراه است؛ برخی از این مشکلات توسط قیومی و همکاران (۲۰۱۳) مورد بررسی قرار گرفته است. آن‌ها معتقدند منبع مشکلات مربوط به عوامل متعددی می‌شود، از جمله: نظام نوشتاری فارسی، تلفیق نظام نوشتاری فارسی با خط عربی، شیوه تایپ کردن تاییست‌ها و خلاقیت زبانی. این مشکلات به صورت خودکار و از طریق برنامه‌نویسی یا به صورت دستی حل شده است. به عنوان مثال، درباره کدهای مربوط به حروف، از همان کدهای عربی استفاده شده است و برخی از کدها که در عربی موجود نیست (مانند کد مربوط به حرف «گ» یا «ژ») به کدهای قبلی اضافه شده است. فاصله‌گذاری درست بین ریشه و وندها که معمولاً تاییست‌ها رعایت نمی‌کنند، خود باعث مشکل در پردازش متن می‌شود؛ به عنوان مثال کلمات «کتاب‌ها»، «کتاب‌ها» و «کتابها» یک کلمه هستند، اما اگر بین «کتاب» و «ها» یک فاصله (و نه نیم‌فاصله) قرار گیرد، در پردازش، دو کلمه در نظر گرفته می‌شوند. این مشکل به این صورت حل شده است که شکل‌های مختلف یک کلمه که در نتیجه اتصال وند به ریشه/ستاک ایجاد شده‌اند، یک شکل در نظر گرفته شوند و همگی با نیم‌فاصله فرض شوند؛ این عمل نیز به صورت خودکار و از طریق برنامه‌نویسی انجام شده است. اما مشکل مربوط به اجزای کلمات مرکب

که معمولاً با ساخت اضافه به هم متصل می‌شوند، را نمی‌توان از طریق مذکور حل کرد. زیرا وندها محدود هستند و در برنامه‌نویسی می‌توان وندهای تصریفی و اشتقاقی را لحاظ کرد تا هر جا به ستاک/ریشه متصل شده باشند، نیم‌فاصله در نظر گرفته شود، اما در مورد کلمات مرکب چون نمی‌توان تعداد ثابتی برای اجزای آن‌ها در نظر گرفت، نمی‌توان چنین برنامه‌نویسی‌ای انجام داد. در این حالت باید شکل-های مختلف کلمه توسط تحلیلگر ساخت‌واژی تشخیص داده شوند. بی‌جن‌خان و همکاران (۲۰۱۱) نیز به بررسی مسائلی می‌پردازند که طی ساخت یک منبع مکتوب زبانی، تحت عنوان «پیکره»، مورد توجه قرار گرفته است. این پیکره حاوی ۳۵/۰۵۸ فایل متنی است که هر کدام حاوی متن کامل یا نمونه تصادفی از یک فایل متنی هستند. اندازه هر متن زنجیره‌ای است با حداقل ۱۰۰۰ کلمه. برای حاشیه‌نویسی فارسی در پیکره، از دستورالعمل EAGLES استفاده شده است. این دستورالعمل برای زبان‌های اروپایی ساخته شده است؛ البته برای زبان‌های غیراروپایی مانند عربی هم استفاده شده است. از این دستورالعمل برای فارسی نیز می‌توان استفاده کرد؛ زیرا علاوه بر اینکه فارسی هم جزء زبان‌های هندواروپایی است، در مقایسه با انگلیسی، ساخت‌واژه تصریفی فارسی در مورد افعال و اسامی به حد کافی غنی است و می‌توان از این دستورالعمل استفاده کرد. نتیجه به دست آمده از به کار گرفتن این دستورالعمل، به صورت سلسله‌مراتبی از برچسب‌ها می‌باشد. جهت حاشیه‌نویسی از روش نیمه‌خودکار استفاده شده است. در تهیه این پیکره، به ساخت اضافه و هم‌نگاره‌ها که در تجزیه و تحلیل متن مشکلاتی ایجاد می‌کنند، توجه خاصی شده است. حسین‌نژاد و همکاران (۱۳۹۶) در این مقاله نحوه و مراحل تهیه پیکره اعلام- یک پیکره استاندارد با برچسب واحدهای اسمی برای زبان فارسی- را شرح می‌دهند. مجموعه تهیه شده با داشتن ۱۳ برچسب واحدهای اسمی و حجم ۲۵۰۰۰۰ کلمه، نیاز سامانه‌های برچسب‌گذاری خودکار در حوزه پردازش زبان طبیعی فارسی را برآورده می‌کند. با استفاده از این پیکره و به کارگیری روش یادگیری ماشینی میدان تصادفی شرطی، سامانه‌ای برای شناسایی واحدهای اسمی جملات فارسی تهیه شده است که دارای دقت ۹۲/۹۴ درصد و فراخوانی ۷۸/۴۸ درصد است.

یکی دیگر از انواع پیکره، پیکره موازی است. پیکره‌های موازی از متونی تشکیل شده‌اند که با ترجمه‌هایی از یک یا بیش از یک زبان دیگر در ارتباط قرار گرفته‌اند. تابع بردبار (۱۳۹۳) به توضیح ایجاد پیکره زبانی موازی به وسیله پیکره‌های قیاس‌پذیر می‌پردازد. به منظور تشکیل این پیکره زبانی، وی از پایگاه دانش ویکی‌پدیا استفاده کرده است. در این پروژه یک شیوه مبتنی بر بازیابی اطلاعات که بر اساس برچسب گروه و لینک کار می‌کند به منظور دسته‌بندی مقالات مشابه استفاده شده است. پس از دسته‌بندی

مقالات مشابه، تعدادی خصیصه که دربرگیرنده ویژگی‌هایی مانند: شباهت ترجمه و میزان شباهت از جهت ترازبندی می‌باشند، برای امتیازدهی جملات استفاده شده است. به منظور دادن وزن بهینه به هر یک از این خصیصه‌ها، از یک مدل خطی استفاده شده است. قاسمی‌زاده و همکاران (۱۳۸۵) نیز پیکره فارسی موازی با تعداد زیادی از زبان‌های اروپایی را معرفی می‌نمایند. آن‌ها در این مقاله، اولین قدم‌ها را برای ساخت منابع اساسی جهت پردازش زبان فارسی معرفی می‌نمایند. این مرحله شامل معرفی ویژگی‌های صرفی-نحوی زبان فارسی و رمزنگاری آن‌ها بر پایه مدل EAGLES/MULTEXT و منابع خاص MULTEXT-East است. در این مقاله پس از معرفی مختصر زبان فارسی، با تاکید بر شیوه نگارش و ویژگی‌های صرفی-نحوی آن، دسته‌بندی جدیدی برای مقوله‌های گفتاری فارسی ارائه می‌شود و رسم-الخطی را جهت ارائه فارسی در محیط دیجیتال ارائه می‌شود. پس از آن آماری از پیکره ساخته شده نیز ارائه می‌شود. حسینی معصوم و غضنفری (۱۳۹۱) بر این باور هستند که ترجمه برخط تحت شبکه یکی از زمینه‌های نوینی است که در حال حاضر دوران ابتدایی تکامل خود را طی می‌کند. مهم‌ترین ویژگی چنین روشی در ترجمه، استفاده از پیکره‌های زبانی پویاست که دارای اساسی تعاملی و آمار-بنیاد است و به طور لحظه‌ای در حال ارتقاء می‌باشد. این مقاله با تکیه بر یافته‌های زبان‌شناسی پیکره‌ای، تقسیم‌بندی جدیدی از انواع پیکره ارائه می‌دهد و ضمن معرفی شیوه عملکرد ترجمه‌های اینترنتی مبتنی بر پیکره‌های زبانی پویا، به معرفی و آزمون عوامل گوناگون، بالاخص عوامل فرهنگی موثر در بروز خطا در ترجمه اینترنتی از فارسی به انگلیسی می‌پردازد. گرچه یافته‌های این تحقیق حاکی از وجود مشکلات جدی در این مسیر است، به نظر می‌رسد الگوریتم‌های پیشرفته نوین در نرم افزارهای معیار ترجمه بتواند گامی امیدوارکننده بسوی تکامل این فناوری و افزایش سرعت و بهبود کیفیت این شیوه از ترجمه ماشینی باشد.

از جمله تلاش‌هایی که در زمینه ساخت پیکره‌های دوزبانه فارسی-انگلیسی انجام گرفته است، می‌توان به پیکره ساخته شده توسط دشتبانی (۱۳۹۱) اشاره کرد. وی اقدام به طراحی و ساخت پیکره دو زبانه فارسی-انگلیسی حوزه فاوا (حوزه فناوری ارتباطات و اطلاعات) کرده است. این پیکره به صورت خودکار ساخته شده است و منابع آن، اسناد تخصصی حوزه فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. علاوه بر این، نرم‌افزار ارائه شده قابلیت مدیریت پیکره را برای کاربران فراهم می‌کند. از ویژگی‌های پیکره ساخته شده، فراهم کردن مجموعه‌ای متمرکز از اسناد تخصصی است که می‌تواند در پروژه‌های مختلف حوزه فاوا استفاده شود. مهم‌ترین مرحله ساخت پیکره‌های چندزبانی، ترازبندی داده‌های پیکره است. در این پروژه

روشی برای ترازبندی جمله‌های پیکره فارسی تخصصی حوزه فاوا و جملات انگلیسی پیکره تخصصی حوزه فاوا ارائه شده است. هدف این پژوهش، طراحی یک سیستم ترازبندی برای استخراج جمله‌های متناظر دو زبان است. در این روش با استفاده از یک لغت‌نامه دوزبانی و تکنیک‌های هوش مصنوعی، امتیاز نشان دهنده شباهت دو جمله، محاسبه می‌شود. در نهایت اطلاعات مربوط به نگاشت جمله‌های دو مجموعه انگلیسی و فارسی در پایگاه داده پیکره، ذخیره می‌گردد.

مرادی و همکاران (۱۳۹۱) اولین پیکره فارسی نظرات برچسب‌خورده از نظر قطبیت (مثبت و منفی بودن) برای کاربردهای تحلیل نظر کاربر و نیز پیکره برچسب‌خورده از لحاظ جنسیت برای تشخیص جنسیت نویسنده متن، بصورت خودکار از اسناد html بدست آمده از صفحات اینترنت ساخته شده است. ایده به کار برده شده در این تحقیق، استفاده از tag های html دارای میزان رضایت و نام نویسنده در صفحه نظرات بوده است. کاربرد این دو پیکره جمع‌آوری شده در این تحقیق، استفاده در طبقه‌بندی و تحلیل متون است.

مدرس و قیومی (۱۳۸۵) با تکیه بر معیار آماری در مطالعه باهم‌آیی این مسئله را مورد بررسی قرار می‌دهند که چگونه دو پیکره زبانی نسبتاً مشابه، به لحاظ نوع و بسامد ساختارهای باهم‌آیی متفاوتند؛ لازمه این کار تهیه کردن دو پیکره زبانی و مقایسه پربسامدترین ساختارهای باهم‌آیند است. سپس آن‌ها به این نتیجه می‌رسند که اگر نمونه‌ای در یک پیکره از میزان باهم‌آیی زیاد برخوردار باشد، به احتمال بسیار زیاد در پیکره دیگر نیز از باهم‌آیی زیاد برخوردار خواهد بود. دلیل تفاوت در میزان باهم‌آیی‌های مشابه در پیکره‌های متفاوت در تفاوت‌های سبکی و توجه خاص به برخی نمونه‌هاست. مدرس خیابانی (۱۳۸۴) سه معیار آماری، معنایی و معنایی - نحوی را در شناخت باهم‌آیی مطرح می‌کند. همانجا به نقل از استابز (۱۹۹۵) به این مسئله اشاره می‌کند که شم زبانی گویشوران بومی یک زبان منبعی قابل اعتماد برای بررسی باهم‌آیی نیست. زیرا گویشوران اغلب تعداد محدودی از نمونه‌های باهم‌آیی را ارائه می‌کنند و ممکن است به شکلی تقریبی درباره احتمال هم‌آیند بودن نمونه‌های خود قضاوت نمایند. از سوی دیگر گویشوران بومی نمی‌توانند حجم وسیعی از باهم‌آیی‌های مستند و میزان بسامد و توزیع آن‌ها را مشخص کنند. بنابراین توجه به پیکره زبانی در مطالعه باهم‌آیی نقش بسزایی خواهد داشت.

از دیگر کاربردهای پیکره، استفاده از پیکره در فرهنگ‌نگاری و نوشتن دستور زبان است. بی‌جن‌خان (۱۳۸۳) به نقش پیکره‌های زبانی در نوشتن دستور زبان پرداخته است. وی معتقد است که اگرچه با تأکید زبان‌شناسان زایشی بر داده‌های بالقوه و نه بالفعل زبانی برای کشف پیچیدگی‌های ذهن انسان، جایگاه

پیکره‌ها در تجزیه و تحلیل دستوری تضعیف شده، اما با رشد سریع فن‌شناسی اطلاعات، ضرورت تهیه پیکره‌ها و دادگان‌های زبانی با حجم بسیار بالا از اولویت بالایی برخوردار شده است. وی پس از تعریف پیکره زبانی که از کریستال (۱۹۹۴) نقل قول می‌کند، به بررسی نقدهایی که بر پیکره‌های زبانی شده است می‌پردازد. سپس به رابطه دستور زبان و پیکره زبانی می‌پردازد. و در نهایت یک نرم‌افزار (نرم‌افزار پیکره زبانی) را معرفی و توصیف می‌کند. مراحل تولید پیکره زبانی عبارتند از: ۱- جمع‌آوری داده‌ها ۲- آماده‌سازی داده‌ها ۳- برچسب‌دهی نحوی - معنایی ۴- آمارگان . وی نحوه جستجوی هوشمند در پیکره زبانی را نیز توضیح می‌دهد.

۲-۳- خلاصه فصل

در این فصل به بررسی مطالعاتی پرداخته شد که در حوزه ساخت پیکره انجام شده است. در این پژوهش‌ها، چکیده‌ای از تجربیات محققین این حوزه درباره ساخت پیکره‌ها و چالش‌هایی که هنگام ساخت پیکره‌های متنی یا نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برخی از این چالش‌ها شامل: نوع نظام نوشتاری، تلفیق نظام‌های نوشتاری، شیوه تایپ کردن تایپیست‌ها، خلاقیت زبانی و.... است. برخی دیگر از این پژوهش‌ها شامل توصیف انواع روش‌های ساخت پیکره، نوع پیکره ساخته شده و حاشیه-نویسی‌های استفاده شده در پیکره است.

فصل سوم

مبانی نظری ساخت پیکرهٔ زبانی

۳-۱- مقدمه

در این فصل، به بررسی مبانی نظری چگونگی ساخت پیکره زبانی که خود شامل پرداختن به مهم‌ترین شاخص‌های کلی ساخت پیکره زبانی، فرآیند ساخت پیکره متنی و فرآیند ساخت پیکره گفتاری است، پرداخته می‌شود.

۳-۲- شاخص‌های کلی ساخت پیکره زبانی

خبرگانی که به تجزیه و تحلیل پیکره زبانی می‌پردازند لزوماً در ساخت پیکره/پیکره‌هایی که از آن‌ها استفاده می‌کنند، تبحر ندارند؛ در حقیقت همیشه این خطر متوجه آنان است که ممکن است پیکره‌ای بسازند که حاوی اطلاعاتی باشد که یا از قبل می‌دانند، و بنابراین اطلاعات جدیدی نیست، و یا می‌توانند جزئیات زبان‌شناسی آن را حدس بزنند. حال آنکه شرایط مطلوب این است که پیکره باید طوری طراحی و ساخته شود که الگوهای ارتباطی جوامعی که آن زبان در آن استفاده می‌شود در پیکره منعکس شود. صرف‌نظر از محتوای اسناد مکتوب و گفتاری پیکره، اسناد باید حاصل جمع‌آوری مکالمات مردم و مطالبی باشد که مردم می‌نویسند یا می‌خوانند. بنابراین، محتویات پیکره باید بدون در نظر گرفتن زبان پیکره و بر اساس کارکرد کاربردی زبان در جامعه مورد استفاده باشد. مهم‌ترین شاخص‌های اصلی ساخت پیکره زبانی شامل: نمونه‌گیری^۱، نمایندگی^۲ / نماینده بودن حجم داده‌های زبانی، توازن^۳، اندازه^۴، نوع پیکره: عمومی^۵ یا تخصصی^۶، یک‌دستی^۷ است (سینکلر: ۲۰۰۴). این شاخص‌ها کلی است و در ساخت هر نوع پیکره‌ای، چه نوشتاری/متنی و چه گفتاری/آوایی لحاظ می‌شوند. اما از آنجایی که اکثر پیکره‌ها، پیکره‌های متنی هستند، مثال‌ها و توضیحات مربوط به ساخت پیکره متنی است. در قسمت‌های بعد به بررسی هر کدام از این شاخص‌ها پرداخته می‌شود..

¹ sampling

² representativeness

³ balance

⁴ size

⁵ general corpus

⁶ specialized corpus

⁷ homogeneity

۳-۲-۱- نمونه‌گیری

روش‌های مورد استفاده در زبان‌شناسی پیکره‌ای تا حد زیادی وابسته به روش نمونه‌گیری است. در طراحی کلی پیکره، ملاحظات از قبیل نوع متون مورد استفاده، تعداد متون، انتخاب متون خاص، طول نمونه‌های متون و... وجود دارد که هر کدام مستلزم تصمیم‌گیری درباره نمونه‌گیری است. تصمیماتی که در این زمینه گرفته می‌شود بر مناسب بودن پیکره برای انواع تجزیه و تحلیل زبانی تاثیر می‌گذارد. نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیه پیکره است. بدیهی است که نمی‌توان هیچ محدودیتی برای زبان طبیعی (اعم از اندازه واژگان، محدوده ساخت‌های معنی‌دار، فرآیندهای تکاملی درون‌زبانی و عوامل بیرونی) قائل شد. هیچ پیکره‌ای، صرف‌نظر از این مسئله که چقدر بزرگ باشد یا بادقت طراحی شده باشد، نمی‌تواند دقیقاً منعکس‌کننده ویژگی‌های زبان طبیعی باشد؛ بنابراین، در تهیه هر پیکره، همانند مطالعه هر پدیده نامتناهی، در عین حال که به محدودیت‌های مذکور واقف هستیم، ناچار هستیم نمونه‌گیری کنیم. در حقیقت، چیزی تحت عنوان «تناسب درست» از مجموعه اجزای جمعیت^۱ نامحدود وجود ندارد (در اینجا منظور از جمعیت، زبان یا گونه زبانی است). هدف هر نمونه‌گیری، کاهش حجم مجموعه داده‌ها است. در تهیه پیکره زبانی تلاش بر این است که پیکره تا حد امکان نماینده زبانی باشد که برای تهیه پیکره انتخاب شده است. حال سوال اساسی این است که چگونه می‌توان از یک زبان نمونه‌گیری کرد؟ برای تصمیم‌گیری درباره نمونه‌گیری دو مسئله باید در نظر گرفته شود: ۱- گرایش / سوگیری^۲ ما نسبت به زبان یا گونه‌ای که قرار است از آن نمونه‌گیری کنیم. طراحان پیکره‌های اولیه، مانند «پیکره براون»^۳ و آن‌هایی که بر اساس مدل «براون» ساخته شده بودند، به دنبال ساخت پیکره‌ای نزدیک به زبان معیار^۴ بودند؛ بنابراین، تنها مستندات منتشر شده انتخاب می‌شدند و اکثر گونه‌های زبانی به طور خودکار حذف می‌شدند. بیشتر پیکره‌هایی که امروزه ساخته می‌شوند نیز از همان سیاست پیروی می‌کنند. برخی از پیکره‌ها، متغیری اصلی را به عنوان بخشی از طراحی پیکره در نظر می‌گیرند؛ به عنوان مثال، یک پیکره تاریخی به لحاظ ساختار درونی، عمداً مقابله‌ای طراحی شده است و هدف آن ارائه تصویری واحد و منسجم از زبان طی زمان نیست. نوع دیگر پیکره که بعد زمان را نیز لحاظ کرده است، پیکره نظارتی^۵ است که نمونه‌های زبانی را در بازه منظم جمع‌آوری می‌کند و نرم‌افزار آن،

^۱ Population

^۲ orientation

^۳ Brown corpus

^۴ standard language

^۵ monitor corpus

تغییرات واژگان و عبارات را ثبت می‌کند. پیکره‌های موازی^۱ (هر پیکره‌ای که از بیش از یک زبان استفاده می‌کند)، حاوی اجزای درونی مقابله‌ای هستند که به تفاوت‌ها و شباهت‌های زبانی دو زبان یا بیشتر در سطوح مختلف می‌پردازند؛ این پیکره‌ها را می‌توان پیکره‌های مقابله‌ای^۲ نامید، انگیزه اصلی ساخت چنین پیکره‌هایی، تقابل اجزای اصلی زبان‌های مورد مطالعه آن‌ها است. بنابراین، سوگیری طراحان پیکره و اینکه به دنبال ساخت چه نوع پیکره‌ای با تقابل قرار دادن چه اجزایی باشند، محتوای پیکره‌ها و نمونه‌گیری را تعیین می‌کند. ۲- معیارهایی که بر اساس آن‌ها نمونه‌ها انتخاب می‌شوند. برخی از معیارهایی که بر اساس آن‌ها نمونه‌گیری صورت می‌پذیرد شامل: شکل^۳ متن (گفتاری / نوشتاری / الکترونیکی)، نوع متن (کتاب / مجله / حتی نامه)، حوزه متن (آکادمیک / عمومی)، زبان (زبان‌ها یا گونه‌های زبانی پیکره)، مکان متن (به عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا). همه معیارهای مذکور از پیش تعیین شده هستند و معمولاً در نام پیکره منعکس می‌شوند. به عنوان مثال، می‌توان از «پیکره انگلیسی گفتاری آکادمیک میشیگان»^۴ نام برد که در دانشگاه میشیگان ساخته شده است و همانطور که از نام پیکره مشخص است، معیارهای استفاده از زبان انگلیسی، گونه گفتاری و آکادمیک در ساخت پیکره لحاظ شده است. اکثر پیکره‌های جامع یک یا چند معیار را که خاص زبان مورد استفاده در پیکره است، برمی‌گزینند. امکان پیش‌بینی دقیق این معیارها در ابتدای امر وجود ندارد. طراح پیکره باید معیارهایی را مقرر کند که آسان به کار برده شوند تا مرتکب خطای کمتری شود؛ زیرا بنانهادهن معیارهای دشوار، پرچالش و پیچیده منجر به خطاهای غیرقابل چشم‌پوشی در طبقه‌بندی و در نهایت بی‌اعتبار کردن پروژه‌های تحقیقاتی عظیم و یافته‌های مربوطه می‌شود. معیارهای تعیین ساختار یک پیکره باید کم و مجزا از هم باشد و برای توصیف پیکره‌ای که نماینده زبان مورد مطالعه است، از کارایی کافی برخوردار باشد. از آنجایی که مسئله توازن و نمایندگی پیکره (که در قسمت‌های بعد در این فصل توضیح داده خواهند شد) وابسته به معیارهای ساختاری است، انتخاب این معیارها از اهمیت بالایی برخوردار است.

نمونه‌های زبانی برای یک پیکره باید تا حد امکان شامل اکثر اسناد مکتوب یا رخدادهای گفتاری مکتوب شده باشد، یا حداقل به هدف مذکور نزدیک باشد. در نمونه‌گیری معمولاً مواردی از قبیل: واحد نمونه‌گیری و محدوده نمونه: کتاب / روزنامه / مقالات / فصول کتاب / مجله، شبکه نمونه: فهرستی از

¹ parallel corpora

² Contrastive corpora

³ mode

⁴ Michigan Corpus of Academic Spoken English

تمام واحدهای ممکن که از میان آن‌ها انتخاب صورت می‌پذیرد. به عنوان مثال، شبکه نمونه «پیکره براون» شامل فهرستی از کتاب‌ها، ماهنامه‌ها و هفته‌نامه‌های موجود در دانشگاه «براون» است (سینکلی: ۲۰۰۴). کاربران پیکره، روش نمونه‌گیری، اندازه نمونه: متن کامل / بریده‌هایی از متون، و ژانر متن در نظر گرفته می‌شود. نمونه‌ها معمولاً بر اساس ژانرهایی مانند اثر منشور داستانی و غیرداستانی، اثر شعری از شاعران معاصر، مجله و نشریه علمی و ادبی و تخصصی، نمایش‌نامه، فیلم‌نامه، ادبیات کودکان، روزنامه و نشریه خبری، همه‌پسند و متنوع، کتاب‌های درسی دبستانی و راهنمایی و دبیرستان، کتاب‌های تالیفی در زمینه آموزش زبان به غیرفارسی‌زبانان، مجموعه‌ای از قوانین و مقررات، دفترچه راهنما و بروشورها، انتخاب می‌شوند.

روش نمونه‌گیری معمولاً به دو صورت «نمونه‌گیری تصادفی ساده»^۱ و «نمونه‌گیری لایه‌لایه»^۲ است. نمونه‌گیری تصادفی ساده، روشی است آماری که در آن همه واحدهای نمونه، داخل شبکه نمونه قرار داده و شماره‌گذاری می‌شوند. در این رویکرد اعضای جمعیت به صورت تصادفی انتخاب می‌شوند و احتمال انتخاب هر عضو از جمعیت از طریق فرمول زیر محاسبه می‌شود (واتام: ۲۰۱۵):

اندازه نمونه

اندازه جمعیت

در نمونه‌گیری تصادفی، همه متون موجود در جمعیت، از شانس یکسانی برای انتخاب شدن برخوردار هستند. این روش مرتکب خطاهایی که ریشه در طبقه‌بندی دارد، نمی‌شود. در نمونه‌گیری لایه‌لایه، حدود متونی که به عنوان نمونه استفاده خواهند شد از قبل تعیین می‌شود، داده‌ها به گروه‌های همگن تقسیم می‌شوند و درباره نسبت استفاده از هر کدام تصمیم‌گیری می‌شود و در نهایت از هر لایه به طور تصادفی انتخاب‌هایی صورت می‌گیرد. در نمونه‌گیری لایه‌لایه، در واقع یک نمونه تصادفی به مجموعه شبکه‌هایی با اندازه‌های خاص، شکسته می‌شود. این روش مستلزم توانایی تقسیم کردن مجموعه به طبقات خاص متقابل و سپس، نمونه‌گیری تصادفی از این طبقات است (واتام: ۲۰۱۵). پارامترهایی که برای این نوع نمونه‌گیری در نظر گرفته می‌شود شامل: کانال اولیه: نوشتاری / گفتاری / گفتار مکتوب شده، شکل: منتشر

^۱ Simple Random Sampling (SRS)

^۲ Stratified sampling

شده/ منتشر نشده، مکان: موسسه/ اماکن عمومی/ اماکن خصوصی، مخاطب: جمع/ فرد/ عام/ خاص، وابستگی به واقعیت: اطلاعات واقعی/ تخیلی، هدف: ترغیب کردن/ سرگرم کردن....، و موضوع است.

این نوع نمونه‌گیری معمولاً نماینده بهتری از زبان مورد مطالعه در نظر گرفته می‌شود. این روش امکان تعریف و آزمودن لایه‌ها را قبل از بازیابی متن فراهم می‌کند.

معیارهای دیگر با نام‌های دیگر و کلی‌تری را نیز می‌توان برای نمونه‌گیری در نظر گرفت. به عنوان مثال در ساخت «پیکره اسکاتلندی»^۱، معیارهایی مانند حوزه^۲، زمان، وسیله بیان^۳ و تناسب هدف^۴، در نظر گرفته شده است و تعاریفی به صورت زیر برای آن‌ها ارائه شده است (اندرسون و همکاران: ۲۰۰۷):

➤ معیار حوزه: حوزه یک متن مربوط به نوع نوشته‌های درون متن است. به عنوان مثال، در «پیکره اسکاتلندی»، ۷۵٪ متون مکتوب حاوی اطلاعات آموزنده در زمینه‌های علوم کاربردی، هنر، تجارت، عقاید و افکار، علوم اجتماعی، علوم محض و هستند. ۲۵٪ هم تخیلی شامل اثرهای ادبی و خلاق هستند.

➤ معیار وسیله بیان: این معیار شامل نوع انتشار متون می‌شود. به عنوان مثال، در «پیکره اسکاتلندی»، ۶۰٪ متون مکتوب پیکره، کتاب است. ۲۵٪ روزنامه و مجلات. ۵ تا ۱۰ درصد، از منابع دیگر منتشر شده. ۵ تا ۱۰ درصد از منابع منتشر نشده مانند نامه‌های خصوصی، دفتر خاطرات.... و در نهایت، ۵٪ هم از مطالبی که برای استفاده در گفتار، نوشته شده‌اند استفاده شده است.

➤ معیار زمان: این معیار مربوط به تاریخ انتشار متون است. به عنوان مثال، اگر پیکره هم‌زمانی باشد، باید متون یک دوره ثبت شده باشد؛ به عنوان مثال، در «پیکره ملی بریتانیا»^۵ نباید متنی قبل‌تر از سال ۱۹۷۵ در پیکره وارد شود.

به عنوان مثال، پیکره بی‌جن‌خان حاوی متونی با مشخصاتی است که در جدول ۳-۱ قابل مشاهده است:

¹ Scottish corpus

² domain

³ medium

⁴ target proportion

⁵ British National Corpus

جدول ۳-۱: مشخصات متون پیکره بی‌جن‌خان (برگرفته از بی‌جن‌خان و همکاران: ۲۰۱۱)

Register	Language Variety	Mode	Factuality	Medium	Sample Size (word)
Novel	SupS, InF	WR	Fiction	Book	11,009,000
Dramas (1)	S, F	WS	Fiction	Book	6,877,000
Academic	SupS, F	WR	Non-Fiction	Book	12,621,000
Newspaper	S/SupS/SubS, F/InF	WR	Non- / Fiction	Newspaper	26,440,000
Popular, general	S, F	WR	Non- / Fiction	Magazine	10,420,000
Periodicals	SupS, F	WR	Non-Fiction	Book	12,380,000
Education	S, F	WR	Non-Fiction	Book	2,549,000
Manuals	S, F	WR	Non-Fiction	Book	930,000
Constitutions	S, F	WR	Non-Fiction	CD	425,000
Government Projects	S, F	WR	Non-Fiction	Web page	335,000
Regulations	S, F	WR	Non-Fiction	CD	1,140,000
Travel Report	S, InF	WR	Non-Fiction	Personal/web page	525,000
Year book	S, F	WR	Non-Fiction	Book	417,000
Correspondences	S, F	WR	Non-Fiction	Archive	84,000
Minutes	S, F	WR	Non-Fiction	Archive	23,000
Daily notes	S, InF	WR	Non-Fiction	Personal	85,000
Presonal letters	S, InF	WR	Non-Fiction	Personal	65,000
Brochures	S, F	WR	Non-Fiction	Web page	1,140,000
Lectures, non-prepared	S, InF	WR	Non-Fiction	Radio/TV	1,600,000
Lectures, prepared	S, F	WS	Non-Fiction	Radio/TV	200,000
Drama (2)	S, InF	WS	Fiction	Radio	5,000,000
Broadcasting news	S, F/InF	WS	Non-Fiction	Radio/TV	785,000
Conversation	SubS, InF	WS	Non-Fiction	Telephone	735,000
Interviews	S, InF	WS	Non-Fiction	Radio/TV	2,900,000

در این جدول اختصاراتی ذکر شده است که به شرح زیر می‌باشد:

S= Standard

SupS= Super-Standard

InF=Informal

F= Formal

SubS=Sub-Standard

WR= Written to be Read

WS= written to be Spoken

معیارهای بسیار دیگری، غیر از آنچه تاکنون مطرح شد، برای نمونه‌گیری در نظر گرفته می‌شود، مانند موضوع متن، نام نویسنده، ژانر، منطقه اقتباس، جنسیت نویسنده، سطح نوشتار و....

برای طراحی نمونه باید مراحلی مانند: تعریف پرسش پژوهش (به منظور تعریف فرضیه‌ها و طراحی نحوه تجزیه و تحلیل؛ هر دو مورد روی اندازه و شکل نمونه تاثیر می‌گذارند)، مشخص کردن واحد نمونه‌گیری، انتخاب متغیرها، تعریف جمعیت (در اینجا زبان یا گونه زبانی مورد مطالعه)، سیاست نمونه‌گیری و اندازه نمونه، در نظر گرفته شوند (واتام: ۲۰۱۵).

۳-۲-۲- نمایندگی

سوال مطرح در طراحی هر پیکره زبانی این است که پیکره زبانی چه ویژگی‌هایی باید داشته باشد تا بتوان آن را نماینده زبان تلقی کرد؟ زمانی می‌توان گفت یک پیکره نماینده یک گونه زبانی است که یافته‌های به دست آمده از تجزیه و تحلیل محتوای پیکره، قابل تعمیم دادن به کل گونه زبانی باشد. به عنوان مثال، فرض کنید بیشترین سند مکتوبی که در بافت ایران مورد استفاده قرار می‌گیرد، روزنامه باشد و از میان روزنامه‌ها، روزنامه همشهری دارای بیشترین خواننده باشد. در این صورت آیا در تهیه پیکره فارسی باید بیشتر متون مستخرج از روزنامه همشهری باشند؟ عمده‌ترین چالش چنین انتخاب‌هایی، مسئله زبان است؛ به این معنی که احتمالاً گونه زبانی استفاده شده در روزنامه‌ها برای گزارش اخبار روز است و استفاده از این گونه به عنوان مدل کلی نوشتار توصیه نمی‌شود. این مثال اهمیت انتخاب نمونه زبانی که گونه زبان مورد استفاده در یک جامعه را نمایندگی کند، نشان می‌دهد. پیکره باید نماینده گونه زبانی باشد تا به عنوان پایه و اساس تعمیم‌های زبانی مورد استفاده قرار گیرد. به عنوان مثال، پایه و اساس فرهنگ‌های لغت پیکره-بنیاد، سیستم‌های برچسب‌دهی به اجزای کلام، دستور زبان و.... مسئله نمایندگی متون استفاده شده برای تهیه آن‌ها است. نمایندگی اکثر پیکره‌ها با دو عامل مشخص می‌شود: ۱- میزان ژانرها/انواع متون موجود در پیکره و ۲- چطور بریده‌های متون برای هر ژانر انتخاب می‌شوند. ارزیابی مسئله نمایندگی یک پیکره بستگی به چگونگی تعریف جمعیت (زبان یا گونه زبانی) مورد استفاده برای نمونه‌گیری دارد. در تعریف جمعیت (زبان یا گونه زبانی)، دو جنبه لحاظ می‌شود: ۱- حد و اندازه جمعیت - چه متونی انتخاب و چه متونی حذف شوند؛ به عنوان مثال، در پیکره براون حد و اندازه به این صورت تعریف می‌شود: همه متون انگلیسی منتشر شده در سال ۱۹۶۱. و ۲- سازماندهی سلسله‌مراتبی

درون جمعیت - چه طبقه‌بندی‌های مربوط به متون در گونه زبانی مورد مطالعه در نظر گرفته شوند؛ به عنوان مثال، در پیکره براون این سلسله‌مراتب به این صورت تعریف می‌شود: ۱۵ طبقه اصلی از متون و تعداد بسیار زیادی زیرمجموعه این طبقات به تفکیک ژانرهای مربوط به هر متن (کواسرا: ۱۹۶۴، نقل از بایبر: ۱۹۹۳).

نمایندگی یک پیکره وابسته به توزیع عناصر درون‌زبانی موجود در جمعیت زبانی نیز هست؛ مشخصه‌های زبانی گوناگون به شکل‌های مختلف درون متن، بین متون و بین ژانرهای متون، توزیع می‌شوند و پیکره نماینده امکان تجزیه و تحلیل انواع گوناگون توزیع زبانی را فراهم می‌آورد. پژوهشگران معمولاً اندازه نمونه را به عنوان مهم‌ترین مسئله برای نمایندگی پیکره مطرح می‌کنند؛ منظور از اندازه نمونه، تعداد متون موجود در پیکره و تعداد کلمات هر نمونه است. به منظور انتخاب متون نماینده، مراحل زیر توصیه می‌شود (سینکلر: ۲۰۰۴):

- تصمیم‌گیری درباره معیارهای ساختاری برای ساخت پیکره و سپس، به کار بردن آن معیارها برای تعیین چارچوب اصلی اجزای پیکره
 - طراحی فهرست جامعی از انواع متون مربوطه که می‌توان برای هر جزء پیکره در نظر گرفت
 - اولویت‌بندی انواع متون با در نظر گرفتن همه عواملی که می‌تواند اهمیت نوع متن انتخابی را کم یا زیاد کند
 - تخمین نهایت اندازه نوع هر متن
 - مقایسه مداوم ابعاد واقعی مطالب و هدف اولیه هم‌زمان با شکل‌گیری پیکره
 - (از همه مهم‌تر) مستند کردن مراحل مذکور به منظور دسترسی کاربران به ارجاع در صورت برخورد به نتایج غیرقابل پیش‌بینی و در نهایت، پایه و اساس قرار دادن تجربه برای پیشبرد پروژه
- به طور کلی، سوال پژوهش تعیین‌کننده میزان نماینده بودن پیکره است. به عنوان مثال، اگر کسی می‌خواهد پیکره‌ای بسازد که نماینده فارسی عمومی باشد، استفاده از متون روزنامه‌ها مناسب نیست؛ همین‌طور اگر کسی می‌خواهد پیکره‌ای بسازد که نماینده روزنامه‌های انگلیسی باشد، نمی‌تواند تنها از روزنامه Times استفاده کند.

۳-۲-۳- توازن

پیش‌نیاز نمایندگی، توازن است. یکی از عواملی که توازن یک پیکره را تعیین می‌کند، میزان استفاده از گونه گفتاری زبان است؛ اکثر پیکره‌ها از این نظر، از توازن خوب و مناسبی برخوردار نیستند، زیرا به اندازه کافی حاوی گونه گفتاری زبان نیستند؛ این در حالی است که مردم بیشتر از گونه گفتاری زبان استفاده می‌کنند تا گونه نوشتاری. بنابراین، در حالت بهینه میزان استفاده از گونه گفتاری در یک پیکره عمومی باید ۵۰ تا ۹۰ درصد باشد. البته باید این نکته را هم مدنظر قرار داد که اجزای نوشتاری یک پیکره زیرشاخه‌هایی مانند روزنامه، مجله، کتاب و... دارد که برای هر کدام لزوماً معادل گونه گفتاری زبان (مانند رسانه‌ها، جلسات، مکالمات و...) وجود ندارد. بنابراین حالت بهینه هیچ‌گاه فراهم نمی‌شود. عامل دیگر تاثیرگذار بر مسئله توازن پیکره، میزان تخصیصی بودن متون پیکره است؛ استفاده از متون بسیار تخصصی در پیکره عمومی باعث بر هم خوردن توازن پیکره می‌شود. بنابراین، نوع متن از اهمیت بالایی برخوردار است (البته در قسمت‌های قبل، نمونه‌گیری و نمایندگی، نیز درباره اهمیت نوع متن صحبت شد). تهیه پیکره بسیار تخصصی مستلزم استفاده از متون بسیار تخصصی مربوط به هر حوزه است و این مسئله توازن را بر هم نمی‌زند و تهیه چنین پیکره‌هایی مسیری کمی متفاوت از تهیه پیکره‌های عمومی را طی می‌کنند.

به طور کلی، دو مسئله «نمایندگی» و «توازن» تعریف خیلی دقیقی ندارند، با این وجود باید در طراحی پیکره و انتخاب اجزای پیکره لحاظ شوند (سینکлер: ۲۰۰۴). به عنوان مثال، چطور یک مجله مربوط به موسیقی یا کامپیوتر می‌تواند نماینده انواع مجلات مربوط به این حوزه باشد؟ مثالی دیگر را فرض کنید: می‌خواهیم پیکره‌ای از روزنامه‌های ایرانی منتشر شده در یک بازه زمانی خاص را تهیه کنیم، اولین مسئله-ای که به نظر می‌رسد این است که تمام مقالات روزنامه‌ها را جمع‌آوری کنیم و کار ساخت پیکره را شروع کنیم. چالش مربوط به توازن پیکره در این مورد به این صورت است که مقالات یک بخش از روزنامه ممکن است بسیار طولانی‌تر از بخش/بخش‌های دیگر روزنامه باشد. برخی پژوهشگران معتقد هستند می‌توان مشکلاتی از این دست را به این صورت حل کرد که نمونه‌های حاوی ۲۰۰۰ تا ۵۰۰۰ کلمه برای هر متن، تهیه کرد و در پیکره وارد کرد. درباره نوع متون هم معتقد هستند، به طور کلی، تلاش برای ایجاد توازن در پیکره می‌تواند به این گونه باشد که متونی انتخاب شوند که زبان آن‌ها خیلی غیررسمی یا خیلی رسمی نباشد. البته انتخاب نوع متن خود وابسته به هدف پژوهش است. اگر هدف بررسی گونه رسمی زبان در بافت اجتماعی باشد، قطعاً باید از پیکره‌ای استفاده شود که حاوی مطالبی به زبان رسمی

باشد. بنابراین، توازن قابل قبول یک پیکره بر حسب استفاده‌های آن تعریف می‌شود. به عنوان مثال، یک پیکره عمومی (مانند «پیکره ملی بریتانیایی» یا پیکره نوشتاری «براون») که حاوی داده‌های نوشتاری و گفتاری باشد، متوازن در نظر گرفته می‌شود؛ چنین پیکره‌هایی حد وسیعی از انواع متون که نماینده زبان یا گونه زبانی مورد مطالعه هستند را پوشش می‌دهد.

علیرغم اینکه در نظر گرفتن توازن، شرط اساسی طراحی پیکره است، هیچ مقیاس و معیار علمی قابل-اعتمادی برای سنجش توازن یک پیکره وجود ندارد. این مسئله عمدتاً شمی و تخمینی است؛ با این وجود تنها اطمینانی که وجود دارد این است که طبقه‌بندی انواع متون در رسیدن به توازن، از اهمیت بالایی برخوردار است و باعث رده‌بندی متون می‌شود. کنترل توازن پیکره، مسئله‌ای است که می‌تواند پس از ساخت پیکره (یا حداقل در پیکره آزمایشی اولیه) صورت پذیرد که خود از طریق دریافت بازخورد کاربران پیکره انجام می‌شود. استفاده‌کنندگان از پیکره هنگام مطالعه پیکره نقاط قوت پیکره را تحسین می‌کنند و نسبت به ضعف‌های پیکره آگاهی دارند و می‌توانند نظرات خود را درباره نقاط قوت و ضعف پیکره انعکاس دهند و در نهایت کنترل توازن پیکره از این طریق صورت می‌پذیرد (آتکینز و همکاران: ۱۹۹۲). نمی‌توانیم صبر کنیم تا یک پیکره به توازن مطلوب برسد و سپس از آن استفاده کنیم و به این اعتبار، هر گونه تجزیه و تحلیل مبتنی بر پیکره را به دلیل عدم توازن، نامعتبر و نامربوط قلمداد کنیم (آتکینز و همکاران: ۱۹۹۲).

روش معمول برای رسیدن به توازن پیکره این است که طراحان پیکره، یک مدل موجود ساخت پیکره را انتخاب کنند و فرض را بر این قرار دهند که توازن در مدل انتخاب‌شده، لحاظ شده است. به عنوان مثال، «پیکره ملی بریتانیایی» عموماً به عنوان پیکره‌ای متوازن شناخته شده است و مدل این پیکره در ساخت پیکره‌های بسیاری از جمله «پیکره ملی آمریکایی»^۱، «پیکره ملی کره‌ای»^۲، «پیکره ملی لهستانی»^۳ و «پیکره ارجاعی روسی»^۴ دنبال شده است (شاروف: ۲۰۰۳، نقل از مک‌انری و همکاران: ۲۰۰۶). ساختار پیکره ملی بریتانیایی در جدول ۲-۳ قابل مشاهده است:

¹ American National Corpus

² Korean National Corpus

³ Polish National Corpus

⁴ Russian Reference Corpus

جدول ۳-۲: ساختار پیکره ملی بریتانیایی: نوع متنی این پیکره (برگرفته از شاروف: ۲۰۰۳، نقل از مک‌انری و همکاران: ۲۰۰۶)

Domain	%	Date	%	Medium	%
Imaginative	21.91	1960-74	2.26	Book	58.58
Arts	8.08	1975-93	89.23	Periodical	31.08
Belief and thought	3.40	Unclassified	8.49	Misc. published	4.38
Commerce/Finance	7.93			Misc. unpublished	4.00
Leisure	11.13			To-be-spoken	1.52
Natural/pure science	4.18			Unclassified	0.40
Applied science	8.21				
Social science	14.80				
World affairs	18.39				
Unclassified	1.93				

این پیکره تقریباً شامل ۱۰۰ میلیون کلمه است. ۹۰٪ آن‌ها متون نوشتاری است و ۱۰٪ داده‌های مکتوب شده گفتاری است.

مسائل مطرح شده در این بخش بیشتر به پیکره‌های نوشتاری مربوط می‌شود، در ساخت پیکره‌های گفتاری مسائلی از قبیل جنسیت، طبقه اجتماعی و اقتصادی، سن، پیش‌زمینه فرهنگی و قومی شرکت‌کنندگان در ضبط گونه گفتاری، در توازن پیکره لحاظ می‌شود.

۳-۲-۴- اندازه

منظور از اندازه در ساخت پیکره، تعداد کلمات پیکره است. نظرات گوناگونی درباره اندازه پیکره وجود دارد. فرض کلی در رویکرد آماری این است که پیکره هر چه بزرگ‌تر باشد، بهتر است. اما در واقع، هیچ رویکرد نظام‌مندی درباره تاثیر اندازه پیکره و ساختار آن وجود ندارد. اندازه اولین نسل پیکره‌ها به یک میلیون کلمه می‌رسید که در نوع خود در آن دوره با امکانات محدود، پیشرفتی عظیم در حوزه ساخت پیکره محسوب می‌شد. نسل دوم پیکره‌ها به بزرگی چندصد میلیون کلمه بودند. همین باعث می‌شود تصور کنیم هر چه پیکره بزرگ‌تر باشد، آسان‌تر می‌توان به بررسی موارد زبانی نادر در آن‌ها پرداخت (مک‌انری و ویلسون: ۲۰۰۱، نقل از بیانچی: ۲۰۱۲). اندازه پیکره تا حدی به هدف مطالعه مربوط می‌شود. به عنوان مثال، بارش و اورت (۲۰۱۳) معتقدند پیکره‌ای کوچک‌تر اما تمیز و متوازن، برای مطالعاتی مانند بررسی هم‌آیندها در بافت، بهتر از پیکره‌ای بزرگ برگرفته از صفحات وب یا متون

روزنامه است. پیکره‌های تخصصی، به طور کلی، کوچک‌تر از پیکره‌های عمومی هستند. به عنوان مثال، اشکال گوناگون کلمه (تصریفی یا اشتقاقی) در پیکره‌های تخصصی بسیار کمتر از پیکره‌های عمومی است. از طرفی، پیکره‌های بزرگ امکان مطالعه علل اصلی وقوع ساخت‌های واژگانی یا دستوری خاص با فراوانی بالا در زبان را فراهم می‌آورند. به عنوان مثال، مطالعه واژه get در ساخت مجهول انگلیسی، مستلزم بهره‌گیری از پیکره‌ای بزرگ است که بتوان این واژه را در بافت بررسی کرد. برخی از عواملی که در تعیین اندازه پیکره لحاظ می‌شوند، سرعت و کارایی نرم‌افزار و توانایی انسان در کارکردن با مقادیر زیاد داده است؛ هم رایانه ممکن است در پردازش مقادیر زیاد داده ناتوان باشد و هم مغز انسان؛ لحاظ کردن این مسئله ممکن است منجر به ساخت پیکره‌های کوچک یا زیرمجموعه‌سازی از پیکره‌های دیگر شود. مسئله اندازه پیکره از چند منظر حائز اهمیت است: برای مقایسه پیکره با پیکره‌های دیگر، انجام تجزیه و تحلیل‌های کمی و محاسبات آماری و مسئله نمایندگی بودن زبانی پیکره (بیانچی: ۲۰۱۲). در مورد اندازه هم می‌توان، یک پیکره را مبنا قرار داد و اندازه کلمات پیکره را مشابه پیکره مبنا تعریف کرد.

۳-۲-۵- نوع پیکره: عمومی یا تخصصی

هدف تهیه پیکره‌های عمومی، نمایش گونه‌های مختلف و ابعاد گوناگون زبان است. نمونه‌ای از این نوع پیکره‌ها، «پیکره ملی بریتانیایی» است (استون و برنارد: ۱۹۹۷). پیکره‌های عمومی معمولاً بزرگ‌تر از پیکره‌های تخصصی هستند؛ هر چه بزرگ‌تر باشد، بهتر است. به عنوان مثال، پیکره ملی بریتانیایی، حاوی ۱۰۰ میلیون کلمه است و شامل متون متوازن در حوزه‌های گوناگون زبان گفتاری و نوشتاری است. این نوع پیکره‌های بزرگ، غالباً پیکره‌های مرجع خوانده می‌شوند؛ زیرا عمدتاً به عنوان پایه و اساس انواع مقایسه و قضاوت درباره گونه‌های زبانی محسوب می‌شوند و امکان مطالعه مقابله‌ای زبان‌ها یا گونه‌های زبانی را فراهم می‌آورند. پیکره‌های تخصصی حاوی متونی از ژانرهای خاص یا دوره زمانی یا بافت خاص هستند. آن‌ها حاوی نمونه‌هایی از ژانر مورد مطالعه هستند. به عنوان مثال، یک پیکره تخصصی می‌تواند حاوی همه نمایش‌نامه‌های شکسپیر باشد. این پیکره‌ها، به عنوان منابع ارزشمند بررسی و مطالعه حوزه‌ای خاص یا ژانری خاص در نظر گرفته می‌شوند. نمونه‌هایی از این نوع پیکره شامل: «پیکره گفتاری میشیگان» است که حاوی ۱,۷ میلیون کلمه است و در دانشگاه میشیگان تهیه شده است، «پیکره بین‌المللی آموزشی

انگلیسی^۱ که حاوی ۲۰۰۰۰ کلمه است و از مطالبی که زبان‌آموزانی که انگلیسی را به عنوان زبان خارجه یاد می‌گیرند، تهیه شده است. پیکره‌های تخصصی حاوی متون خاص است و هدف آن‌ها نمایندگی زبان آن متون است. این پیکره‌ها عمدتاً برای پاسخ به سوالات بسیار خاص ساخته می‌شوند. به عنوان مثال، یک پیکره پزشکی از زبانی استفاده می‌کند که پرستاران و کارکنان بیمارستان استفاده می‌کنند و حاوی کلمات تخصصی رشته پزشکی و پیراپزشکی است. استفاده از این پیکره‌ها برای اهداف آموزشی نیز هست. به عنوان مثال، پیکره‌هایی برای آموزش زبان تخصصی انگلیسی ساخته می‌شود که به پیکره‌های ESP^۲ معروف هستند. پیکره‌های تخصصی خود می‌توانند مستخرج از یک پیکره عمومی باشند و زیرمجموعه‌ای از پیکره‌های عمومی در نظر گرفته شوند.

۳-۲-۶- یک‌دستی

هیچ مقیاس دقیقی برای اندازه‌گیری میزان یک‌دستی یک پیکره تعریف نشده است. یک پیکره زمانی یک‌دست در نظر گرفته می‌شود که حاوی تفاوت‌های بارز در مشخصه‌های درون مستندات خود نباشد (کاواگلیا: ۲۰۰۲). در هر گونه زبانی که در پیکره مورد استفاده قرار می‌گیرد، متونی یافت می‌شود که با متون دیگر متفاوت هستند و در یک طبقه قرار نمی‌گیرند و در حقیقت نمی‌توانند نماینده گونه زبانی مورد مطالعه باشند. وارد کردن این متون در پیکره، تنها حجم پیکره را افزایش می‌دهد، بدون آنکه کارکرد خاصی داشته باشند و وجودشان در تحلیل‌های زبانی مفید واقع شود؛ به طور کلی، اگر محدودیت حجم داده‌های پیکره مسئله‌ای لحاظ‌شدنی در تهیه پیکره باشد، بهتر است این متون از پیکره حذف شوند یا هنگام ساخت پیکره، وارد نشوند (سینکلر: ۲۰۰۴). شاخص‌های اصلی تهیه هر پیکره زبانی را می‌توان به صورت زیر نمایش داد:

^۱ international corpus of learner English

^۲ English for Special Purposes



شکل ۳-۱: شاخص‌های اصلی تهیه پیکره زبانی

۳-۳- فرآیند ساخت پیکره متنی

در ساخت پیکره متنی، پس از انتخاب متون مربوطه، باید تمهیداتی کلی در نظر گرفته شود که شامل موارد زیر است:

از متونی که در پیکره استفاده خواهند شد، به همان شکلی که از منبع اصلی استفاده شده است، کپی تهیه شود و جهت ارجاع حفظ شود. کپی به صورت چاپ شده باشد و نه الکترونیکی. متن به شکل «متن ساده»^۱ ذخیره شود. در ابتدای هر متن، یک شماره سریال قرار داده شود تا از متون دیگر قابل تشخیص باشد تا مواقع لازم بتوان به آن به راحتی ارجاع داد. البته می‌توان متون را بر اساس مواردی مانند اختصاص نام نویسنده یا تاریخ و.... تفکیک کرد. باید پیش‌پردازش‌هایی روی متن انجام شود (این بخش به تفصیل توضیح داده خواهد شد). حاشیه‌نویسی‌های مربوطه را به صورت دستی یا با استفاده از ابزارهای مربوطه

¹ plain text

انجام داد (این بخش نیز به تفصیل توضیح داده خواهد شد). بعد از اتمام ساخت پیکره، یک کپی از کل پیکره روی CD تهیه شود.

۳-۳-۱- پیش‌پردازش متون

در ساخت پیکره متنی، پس از انتخاب متونی که در پیکره متنی مورد استفاده قرار خواهند گرفت، باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «نرمال‌سازی متن»^۱ نامیده می‌شود. این پیش‌پردازش‌ها در متون فارسی می‌تواند شامل، یک‌دست کردن فاصله‌ها، نشانه‌گذاری‌های درون متن، یکسان کردن کاراکترهای استفاده شده در متون (مانند انواع «ی»، «ک»، همزه و...)، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چنداملائی و یکسان در نظر گرفتن آن‌ها و... باشد. ابتدا باید همه نویسه‌های (کاراکترهای) متن با جایگزینی با معادل استاندارد آن، یکسان‌سازی گردند. متون مختلف ممکن است بسیار به هم شبیه باشند اما به دلیل تفاوت‌های ساده ظاهری از نظر ماشین متفاوت باشند؛ به همین دلیل ابتدا باید این تفاوت‌های ساده ظاهری برطرف گردد. همچنین اصلاحات دیگری نیز به منظور پردازش دقیق‌تر متون در این مرحله صورت می‌شود. برای رسیدن به این هدف، قبل از مقایسه متون، پیش‌پردازش‌هایی روی آن‌ها انجام می‌شود. طبیعتاً هر چه این پیش‌پردازش‌ها قوی‌تر باشد، نتایج حاصل از مقایسه متون قابل اطمینان‌تر خواهد بود. در پردازش رسم‌الخط زبان فارسی، با توجه به قرابتی که با رسم‌الخط عربی دارد، همواره در پردازش تعدادی از نگاره‌ها مشکلاتی وجود دارد که در اولین گام باید مشکلات مربوط به این نگاره‌ها را برطرف ساخت. علاوه بر این، اصلاح و یکسان‌سازی نویسه نیم‌فاصله و فاصله در کاربردهای مختلف آن و همچنین حذف نویسه «ا» که برای کشش نویسه‌های چسبان مورد استفاده قرار می‌گیرد و مواردی مشابه برای یکسان‌سازی متون، از اقدامات لازم قبل از شروع فازهای مختلف می‌باشد. برای انجام پیش‌پردازش‌هایی و نیز انواع حاشیه‌نویسی، می‌توان از ابزارهایی که برای این منظور ساخته شده‌اند استفاده کرد. به طور کلی در جهت تسهیل فرآیند ساخت پیکره و صرفه‌جویی در استفاده از نیروی انسانی و تسریع فرآیند ساخت یا تجزیه و تحلیل پیکره، ابزارهایی تعریف شده است که به برخی از آن‌ها اشاره می‌شود:

¹ normalization

➤ ابزار جداساز^۱: ابزاری برای شکستن یک متن بر اساس واحدهای با معنی مانند کلمه، پاراگراف و نمادهای معنادار مانند space (فاصله) و tab و ... است؛ لازمه ایجاد این ابزار در فارسی، جمع‌آوری واحدهایی است که در زبان فارسی به عنوان واحدهای مستقل معنایی شناخته می‌شوند؛ سپس بر اساس انتخاب هر کدام از این واحدها، متن بر اساس آن شکسته خواهد شد. از نمونه‌های انگلیسی آن می‌توان به Flex، JLex، JFlex، ANTLR، Ragel و Quex اشاره کرد. یکی از مهم‌ترین و پرچالش‌ترین مسائل در این رابطه، تعیین مرز کلمات در متون فارسی است (قیومی و ممتازی: ۲۰۰۹، نقل از بی‌جن‌خان و همکاران: ۲۰۱۱)؛ در فارسی وندها به اشکال گوناگون به ستاک متصل می‌شوند و لزوماً تاییست‌ها از یک رسم‌الخط واحد پیروی نمی‌کنند. به عنوان مثال، برای نشان دادن ترکیب <ه> غیرملفوظ و کسره اضافه، سه شکل ممکن نوشتاری در متون فارسی یافت می‌شود: <ه>، <ه‌ی> و <ه‌ی>. شکل دوم (<ه‌ی>)، توالی است از <ه> و <ی> که برای نگه داشتن شکل <ه> آخر <ه> نویسه نیم‌فاصله ^۲ ZWNJ (U+200c) بعد از <ه> اضافه می‌شود تا از اتصال <ه> به <ی> جلوگیری شود (اسفهد: ۲۰۰۴، نقل از علایی و بی‌جن‌خان: ۱۳۹۲). شکل سوم <ه‌ی> همان توالی است و تفاوت آن با شکل دوم این است که یک فاصله بین <ه> و <ی> درج شده است. به طور کلی می‌توان گفت این سه شکل نوشتاری برای یک دنباله نوشتاری که متشکل از ستاک و وند است به کار می‌رود و در واقع بر حسب شکل نگاره آخر ستاک و اولین نگاره وند ایجاد می‌شوند (مگردومیان ۲۰۰۴). برای مثال برای توالی «کتاب» - که ستاک است - و «ها» - که وند است - سه شکل نوشتاری وجود دارد: «کتابها»، «کتاب‌ها» و «کتاب ها». مثال دیگر ۹ شکل نوشتاری برای «می‌فروخته‌ام» است: «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام»، «می‌فروخته‌ام» (بی‌جن‌خان و همکاران: ۲۰۱۱).

^۱ tokenizer

^۲ Zero-Width Non-Joiner (ZWNJ)

نشانه کنترل‌کننده‌ایست که در محل اتصال تکواژ به ستاک به کار برده می‌شود تا تکواژ و ستاک به هم وصل شوند و یک کلمه نوشتاری

➤ پارسر^۱: تجزیه و تحلیل جمله و شکستن آن به اجزای تشکیل دهنده مانند گروه‌های اسمی، فعلی، قیدی و غیره توسط ابزاری به نام پارسر صورت می‌گیرد که نقش اساسی در طراحی و یا افزایش دقت سایر ابزارهای پردازش متن دارد.

➤ ابزار ریشه‌یابی^۲: ابزاری برای ریشه‌یابی لغات و تشخیص نوع کلمه ساخته شده از آن ریشه (اسم مکان، اسم زمان، صفت فاعلی، مفعولی و...) است. هدف الگوریتم‌های ریشه‌یابی، حذف پیشوندها و پسوندهای کلمات و تعیین ریشه اصلی کلمه است. معروف‌ترین الگوریتم ریشه‌یابی در انگلیسی porter می‌باشد. معمولاً ریشه‌یابی لغات بر اساس قواعد ساخت‌وازی و سپس حذف پیشوندها و پسوندها صورت می‌پذیرد. همچنین در ریشه‌یابی از فرهنگ‌های لغات هم استفاده می‌شود.

➤ ابزار تشخیص گروه‌های نحوی^۳: ابزاری برای تشخیص گروه‌های اسمی، فعلی و در یک جمله است. از جمله نمونه‌های انگلیسی آن می‌توان به Illinois Chunker اشاره کرد.

➤ ابزار برچسب‌دهنده نقش نحوی^۴: ابزاری برای تشخیص نقش نحوی کلمه در جمله است. این ابزار یکی از مهم‌ترین نقش‌ها را در پردازش‌های زبانی بر عهده دارد. دقت در این ابزار بسیار حائز اهمیت است. این ابزار باید نقش‌های نحوی کلمات در جمله‌ها مانند فاعل، مفعول مستقیم، مفعول غیر مستقیم و را تشخیص دهد. از جمله نمونه‌های انگلیسی آن می‌توان به OpenNLP، Illinois SRL و Swirl و LTHSRL اشاره کرد. این ابزارها از الگوریتم پارسینگ charniak استفاده می‌کنند.

➤ ابزار برچسب‌گذاری اجزای واژگانی کلام^۵: ابزاری برای مشخص کردن نوع کلمات از قبیل اسم، صفت، قید، فعل و ... است. به عمل انتساب برچسب‌های واژگانی (اسم، صفت، قید، فعل و ...) به کلمات و نشانه‌های تشکیل‌دهنده یک متن برچسب‌گذاری کلمات گفته می‌شود؛ به صورتی که این برچسب‌ها نشان‌دهنده مقوله کلمات و نشانه‌ها در جمله باشد.

➤ ابزار خطایاب^۶: هدف از توسعه این ابزارها، تشخیص و رفع خطاهای معنایی و نحوی جملات است. خطاهای معنایی و واژگانی به خطاهایی گفته می‌شود که به شیوه املائی و نگارشی لغات

¹ parser

² stemming

³ chunker

⁴ syntactic role labeler

⁵ Part Of Speech (POS) tagging

⁶ error detection

برمی‌گردد. تحلیل نحوی نیز از حوزه‌هایی است که در همه زبان‌های امروز دنیا مورد توجه قرار گرفته است. توجه تنها به واژگان بدون در نظر گرفتن جایگاه دستوری آن‌ها به هیچ وجه برای تحلیل و خطایابی یک متن کافی نیست. به عنوان مثال، جمله «امروز دارد باران می‌بارم» از لحاظ واژگانی کاملاً درست است، ولی از لحاظ نحوی نادرست است. در واقع در جمله اخیر واژه «می‌بارم» یک واژه سردرگم است. واژه سردرگم در واقع کلماتی هستند که از لحاظ املائی درست ولی از لحاظ نحوی در جای نادرستی قرار گرفته‌اند. اگر جمله «سیب مرا خورد» حتی از لحاظ نحوی درست است ولی از لحاظ معنایی نادرست است. با ساز و کارهای موجود در بسیاری از خطایاب‌های نحوی میتوان به صورت محدود به خطایابی معنایی نیز پرداخت. جمله «رئیس جمهور امریکا یک کودک شش ساله است» از لحاظ معنایی نیز درست است ولی در حوزه عمل، چنین چیزی با توجه به قوانین و محدودیت سنی ریاست جمهوری نادرست می‌شود. با وجود اینکه برای خطایابی کامل باید وارد حوزه معنا و کاربرد شد؛ نیاز است که با استفاده از روش‌های مرسوم، از لحاظ نحوی جملات را مورد بررسی قرار داد. این روش‌ها وابسته به دستور زبان خاص و زبان خاصی نیستند (یاری: ۱۳۹۴).

این ابزارها بعضاً برای زبان فارسی نیز ساخته شده‌اند و در صورت چک کردن درصد صحت آن‌ها، می‌توان در ساخت پیکره از آن‌ها استفاده کرد یا تغییراتی را در کدهای مربوطه ایجاد کرد و برای اهداف خاص ساخت پیکره مورد استفاده قرار داد.

اکثر پژوهشگران حوزه پیکره معتقدند جمع‌آوری داده‌ها (چه متنی و چه گفتاری)، به خودی‌خود، ارزش زیادی ندارد و باید اطلاعات گوناگونی که حاوی مشخصه‌های زبانی است، به داده‌ها اضافه شود. بنابراین، یکی از مسائل مهم در ساخت هر پیکره زبانی را حاشیه‌نویسی^۱ می‌دانند.

۳-۲- حاشیه‌نویسی

حاشیه‌نویسی پیکره به عمل اضافه کردن اطلاعات زبانی توضیحی-تفسیری به پیکره اطلاق می‌شود. به عنوان مثال، یکی از رایج‌ترین نوع حاشیه‌نویسی پیکره، اضافه کردن برچسب‌ها به کلمات پیکره است؛ این

¹ annotation

برچسب‌ها می‌تواند نشان‌دهنده مقوله کلمات (اسم، فعل، صفت...) باشد که «برچسب‌گذاری اجزای واژگانی کلام» نامیده می‌شود. یکی از کاربردهای این نوع برچسب‌گذاری، تمایز کلماتی است که یکسان نوشته می‌شوند اما معنی و گاهی تلفظ متفاوت دارند و در اصطلاح «هم‌نگاره»^۱ خوانده می‌شوند. به عنوان مثال، صورت نوشتاری <present> در انگلیسی ممکن است از مقوله «اسم» باشد (به معنی «هدیه»)، «فعل» باشد (به معنی «هدیه دادن») یا «صفت» باشد (به معنی «حاضر»). این صورت نوشتاری می‌تواند در پیکره به صورت زیر برچسب‌گذاری شده باشد:

present_NN1 (singular common noun)
present_VVB (base form of a lexical verb)
present_JJ (general adjective)

برخی از پژوهشگران مانند سینکلر (۲۰۰۴، نقل از لیچ: ۲۰۰۴)، ترجیح می‌دهند وارد مقوله حاشیه-نویسی پیکره نشوند؛ زیرا معتقد هستند پیکره حاشیه‌نویسی نشده، پیکره‌ای خالص است و برای مطالعات زبان‌شناسی چنین پیکره‌هایی را ترجیح می‌دهند. این پژوهشگران به حاشیه‌نویسی اعتماد ندارند و با شک به آن می‌نگرند و معتقد هستند حاشیه‌نویسی نمی‌تواند بدون خطا باشد. اما اکثر پژوهشگران این حوزه بر این باورند که حاشیه‌نویسی، پیکره را بسیار مفیدتر و غنی‌تر از پیکره خام می‌کند؛ از نظر آن‌ها حاشیه-نویسی، ارزشی مضاعف به پیکره می‌دهد. به عنوان مثال، عمل برچسب‌گذاری اجزای واژگانی کلام که برای پیکره برآون انجام شده است، در سطح گسترده‌ای توزیع شده و مورد استفاده افرادی قرار گرفته است که در راستای پژوهش‌های خود نیازمند برچسب‌گذاری اجزای واژگانی کلام یا پیکره برچسب-گذاری شده بودند. غیر از برچسب‌گذاری اجزای واژگانی کلام، حاشیه‌نویسی‌های دیگری نیز وجود دارد که به سطوح مختلف تجزیه و تحلیل‌های زبانی یک پیکره مربوط می‌شود. به برخی از رایج‌ترین آن‌ها اشاره می‌کنیم (لیچ: ۲۰۰۴):

➤ حاشیه‌نویسی آوایی^۲: اضافه کردن اطلاعاتی درباره نحوه تلفظ یک کلمه در پیکره گفتاری یا اضافه کردن مشخصه‌های زبرزنجیری^۳ (مانند تکیه^۴، آهنگ کلام^۵، وقفه^۶ و...) به کلمات تشکیل‌دهنده یک پیکره است.

^۱ homograph

^۲ phonetic annotation

^۳ prosodic

^۴ stress

^۵ intonation

^۶ pause

- حاشیه‌نویسی نحوی^۱: اضافه کردن اطلاعاتی درباره نحوه تجزیه یک جمله به عبارت‌ها و اجزای تشکیل‌دهنده آن است.
- حاشیه‌نویسی معنایی^۲: اضافه کردن اطلاعاتی درباره مقوله / طبقه‌بندی معنایی کلمات پیکره است. به عنوان مثال، صورت نوشتاری <cricket> در انگلیسی، فارغ از صورت نوشتاری یا تلفظی یکسان، هم در طبقه‌بندی انواع ورزش (نوعی ورزش) قرار می‌گیرد و هم در طبقه‌بندی دیگر و متفاوتی تحت عنوان حشرات (نوعی حشره).
- حاشیه‌نویسی کاربردی^۳: اضافه کردن اطلاعاتی درباره کنش گفتار^۴ که در مکالمه اتفاق می‌افتد. به عنوان مثال، پاره گفتار <okay> در انگلیسی، در موقعیت‌های مختلف می‌تواند به منزله اقرار و تصدیق، درخواست بازخورد، پذیرفتن یا نشانه شروع مرحله جدیدی از بحث باشد.
- حاشیه‌نویسی گفتمانی^۵: اضافه کردن اطلاعاتی درباره ارتباطات ارجاعی^۶ در متن است. به عنوان مثال، ارتباط دادن ضمیر <them> و مرجع آن <the horses> در جمله زیر:
I'll saddle the horses and bring them round.
- حاشیه‌نویسی سبکی^۷: اضافه کردن اطلاعاتی درباره نمود گفتار (گفتار مستقیم^۸، گفتار غیرمستقیم^۹...)
- حاشیه‌نویسی واژگانی^{۱۰}: اضافه کردن اطلاعاتی مانند ریشه کلمه.

موارد نام‌برده شده، معروف‌ترین و متداول‌ترین نوع حاشیه‌نویسی هستند؛ می‌توان انواع دیگر حاشیه‌نویسی را از ادبیات مربوطه استخراج کرد یا حتی حاشیه‌نویسی‌های جدیدی را به فهرست حاشیه‌نویسی‌ها اضافه کرد. به عنوان مثال، می‌توان در پیکره‌های آموزشی زبان برچسبی تحت عنوان «برچسب خطا» تعریف کرد که مربوط به خطاهایی است که زبان‌آموزان در بافت‌های گوناگون مرتکب می‌شوند (گرنگر و همکاران: ۲۰۰۲، نقل از لیچ: ۲۰۰۴). برخی از مزایای حاشیه‌نویسی شامل موارد زیر است (لیچ: ۲۰۰۴):

¹ syntactic annotation

² semantic annotation

³ pragmatic annotation

⁴ speech act

⁵ discourse annotation

⁶ anaphoric

⁷ stylistic annotation

⁸ direct speech

⁹ indirect speech

¹⁰ lexical annotation

سنجش سریع محتوای پیکره: اطلاعاتی که به صورت حاشیه‌نویسی در پیکره موجود می‌باشد، قابلیت استخراج مجدد و استفاده‌های گوناگون دارد. به عنوان مثال، یکی از کاربردهای اساسی برچسب‌گذاری اجزای کلام، بالابردن قابلیت و کاربرد پیکره در ساخت فرهنگ‌های زبانی است. فرهنگ‌نگاری که در جستجوی شباهت‌های کلمات و متمایز کردن هم‌نگاره‌ها است (مانند متمایز کردن <separate> «فعل» از <separate> «صفت»)، چنانچه این تمایز از طریق برچسب‌های مربوطه در پیکره مشخص شده باشد، عمل تمایز به صورت خودکار صورت می‌پذیرد و نیازی به جستجوی هزاران مثال برای این منظور نیست. مثال دیگر: اگر دستورنویسی در جستجوی بررسی شکل استمرار افعال باشد، می‌تواند این کار را با ابزار جستجوی موجود در پیکره انجام دهد. تجزیه و تحلیل خودکار پیکره: اگر پیکره‌ای حاشیه‌نویسی شده باشد، امکان بسیاری از پردازش‌های زبانی و تجزیه و تحلیل‌های خودکار فراهم می‌شود. به عنوان مثال، از پیکره‌ای که برچسب اجزای کلام دارد، می‌توان فهرستی از فراوانی کلمات همراه با مقوله نحوی آن‌ها فراهم کرد؛ چنین فهرستی اطلاعاتی از این دست در اختیار ما قرار می‌دهد: صورت نوشتاری <leaves> «فعل» و <leaves> «اسم»، کلمات مجزا هستند و فراوانی آن‌ها، به طور خودکار، مجزا از هم محاسبه می‌شود. از کاربردهای دیگر برچسب‌گذاری، امکان تقطیع نحوی جملات با سرعت و صحت بالاتر است. از برچسب‌گذاری در سنتز گفتار نیز استفاده می‌شود؛ ابزار سنتز گفتار توانایی تشخیص کلمات مشابه را دارد که خود مستلزم استفاده از ابزار برچسب‌گذاری است. استفاده مجدد از حاشیه‌نویسی: به دلیل ماهیت مبهم و پیچیده زبان، حاشیه‌نویسی زبانی نمی‌تواند کاملاً صحیح، بدون خطا و خودکار صورت پذیرد؛ صحت برچسب‌گذاری (که حاشیه‌نویسی نسبتاً ساده محسوب می‌شود) در بهترین حالت به ۹۵ تا ۹۸ درصد می‌رسد که شرایط ایده‌آل نیست؛ بنابراین لازم است برچسب‌ها به صورت دستی نیز کنترل شوند. در واقع، پیکره‌ای که برچسب‌گذاری آن به صورت خودکار انجام گرفته است، باید توسط یک تیم انسانی متخصص نیز ویرایش شود که خود باعث بالاتر رفتن کارایی و ارزش پیکره برای استفاده کاربران می‌شود. زیرا، چنین پیکره‌ای به راحتی در دسترس دیگران (و نه صرفاً کسانی که آن را حاشیه‌نویسی کرده‌اند) قرار می‌گیرد. به طور کلی، پیکره حاشیه‌نویسی شده مرجعی است که همه می‌توانند برای پژوهش‌های گوناگون و مطالعات پیکره-بنیاد در ساخت و بهبود آن سهم شوند. ویژگی چندکارکردی: انواع حاشیه‌نویسی به طور همزمان کاربردهای گوناگونی دارند. به عنوان مثال، همان اطلاعات مربوط به مقوله کلمات می‌تواند همزمان در فرهنگ‌نویسی، تقطیع نحوی

جملات، تهیه فهرست فراوانی واژه‌ها و برچسب‌های آن‌ها، سنتز گفتار و موارد دیگر مورد استفاده قرار گیرد.

دو راه برای حاشیه‌نویسی وجود دارد: ۱- پیروی کردن از استانداردهای خاص که می‌توان آن‌ها را اساس حاشیه‌نویسی پیکره در نظر گرفت؛ ۲- راه دیگر این است که از ابزارهای خاص استفاده کنیم و حاشیه‌نویسی را به کمک نرم‌افزارهای خاصی که برای این منظور طراحی شده است، انجام دهیم (به عنوان مثال، ابزارهایی برای برچسب‌گذاری اجزای کلام یا ویرایش متن وجود دارد که می‌توان از آن‌ها استفاده کرد) و در نهایت، خروجی این حاشیه‌نویسی‌ها را دستی کنترل کنیم.

حاشیه‌نویسی پیکره مستلزم پیروی از استانداردهایی است مانند اینکه: حاشیه‌نویسی باید مجزا از متن پیکره باشد و امکان جدا کردن متن خام بدون حاشیه‌نویسی از متن با حاشیه‌نویسی وجود داشته باشد. باید مستندسازی صریح همراه با جزئیات صورت پذیرد. به عنوان مثال، باید مشخص شود چه کسی، کجا و چگونه حاشیه‌نویسی را انجام داده است؟ چه کدهایی برای این منظور نوشته یا استفاده شده است؟ چه اندازه حاشیه‌نویسی خوب عمل می‌کند و به کار می‌آید؟ عمل حاشیه‌نویسی حتی‌الامکان مورد توافق حاشیه‌نویسان حوزه زبان‌شناسی پیکره‌ای باشد. به عنوان مثال، سر این مسئله که چه چیز را «عبارت فعلی» بنامیم، اتفاق نظر نسبی وجود داشته باشد (لیچ: ۲۰۰۴).

به طور کلی، پس از انتخاب متونی که در پیکره قرار خواهند گرفت (نمونه‌گیری) و در نظر گرفتن معیارهای ذکر شده برای انتخاب نمونه، پیش‌پردازش‌هایی روی متن صورت می‌پذیرد، پس از نرمال‌سازی متن، حاشیه‌نویسی انجام می‌شود؛ در طول ساخت پیکره باید از متخصصین کمک گرفت؛ به عنوان مثال، اگر حاشیه‌نویسی یا نرمال‌سازی مستلزم استفاده از نرم‌افزارهای خاصی است که باید تغییراتی در کدها یا مدل آن‌ها صورت گیرد، برنامه‌نویسان رایانه می‌توانند در این زمینه در ساخت پیکره همکاری داشته باشند؛ همچنین طراحی ساختار و شکل کلی پیکره توسط برنامه‌نویسان متخصص این حوزه انجام می‌شود. اگر حاشیه‌نویسی‌ها دستی انجام می‌شود یا دستی کنترل می‌شود، باید از متخصصان حوزه دستور زبان نیز کمک گرفت. بنابراین ساخت پیکره، فرآیندی تیمی است که هر چه متخصصین با تخصص‌های گوناگون بیشتر همکاری داشته باشند، پیکره‌ای با صحت و کارایی بالاتری ساخته خواهد شد. فرآیند ساخت پیکره متنی را می‌توان به صورت شکل ۳-۲ نمایش داد:



شکل ۳-۲: فرآیند ساخت پیکره متنی

اگر بخواهیم مثالی از نحوه ساخت یک پیکره ارائه دهیم، می‌توانیم به پیکره ملی بریتانیا اشاره کنیم. ساخت پیکره ملی بریتانیا در دو مرحله انجام شده است: ۱- مرحله طراحی و ۲- مرحله ساخت. طی مرحله طراحی، اصولی که برای طراحی پیکره باید مدنظر قرار گیرد، بررسی شده است. این اصول شامل معیارهای انتخاب متون (برای پیکره متنی) است. این معیارها شامل: حوزه، زمان و وسیله بیان است (برنارد: ۲۰۰۷). درباره هر کدام از معیارهای مذکور در بخش‌های پیشین توضیحاتی ارائه شده است و این معیارها در جدول ۳-۲، به صورت کلی ذکر شده‌اند. در این قسمت اطلاعات مربوط به هر بخش مفصلاً به صورت جدول‌های زیر آورده می‌شود: (در این جدول‌ها s-unit= sentence unit و w-unit= word unit است)

جدول ۳-۳: حوزه نوشتاری متون (برنارد: ۲۰۰۷)

	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

جدول ۳-۴: تاریخ انتشار متون (برنارد: ۲۰۰۷)

	texts	w-units	%	s-units	%
Unknown	162	1831585	1.86	126416	2.09
1960-1974	46	1718449	1.74	119510	1.98
1975-1984	169	4730889	4.80	257962	4.28
1985-1993	3672	90082860	91.58	5522396	91.63

جدول ۳-۵: وسیله بیان مربوط به متون نوشتاری (برنارد: ۲۰۰۷)

	texts	w-units	%	s-units	%
Book	1411	50293803	57.18	2887523	57.88
Periodical	1208	28609494	32.52	1487644	29.82
Miscellaneous published	238	4233135	4.81	287700	5.76
Miscellaneous unpublished	249	3538882	4.02	220672	4.42
To-be-spoken	35	1278618	1.45	104665	2.09

پس از طراحی پیکره، مرحله ساخت پیکره آغاز شده است. این مرحله مستلزم انجام مراحل زیر است:

➤ دریافت اجازه‌نامه‌ها از صاحبان متونی که در پیکره استفاده می‌شوند (نویسندگان، ناشران یا متولیان)

➤ جمع‌آوری متون: در ساخت پیکره ملی بریتانیا، سازمان‌های گوناگونی همکاری کرده‌اند. به عنوان مثال، انتشارات دانشگاه آکسفورد مسول تهیه متون نوشتاری پیکره متنی ملی بریتانیا و لانگ‌من، مسول تهیه مواد گفتاری برای پیکره گفتاری ملی بریتانیا بود.

➤ حاشیه‌نویسی زبانی متون: این مرحله شامل برچسب‌گذاری اجزای کلام در هر متن است. همچنین هر متن به اجزاء جمله / شبهه جمله تجزیه شده است. این مراحل با کمک نرم‌افزارهای مربوطه در دانشگاه لنکستر انجام شده است.

➤ ذخیره و تهیه اسناد مربوط به متون: این مرحله شامل اضافه کردن اطلاعات توصیفی مفصل مربوط به هر متن است؛ این اطلاعات شامل نام نویسنده‌های متون یا مکان انجام مکالمه و ضبط آن (در مورد پیکره گفتاری) است.

نکاتی که تا کنون مطرح شده است، نوعاً بیشتر مربوط به ساخت پیکره‌های متنی می‌شود. اما با اینکه تمرکز اصلی این پژوهش، ساخت پیکره متنی است، خوب است ملاحظات کلی ساخت پیکره‌های گفتاری، جهت مقایسه با روند ساخت پیکره‌های متنی، ذکر شود. در این قسمت به بررسی کلی چگونگی ساخت پیکره‌های گفتاری پرداخته می‌شود.

۳-۴- فرآیند ساخت پیکره گفتاری

کار کردن روی داده‌های گفتاری زبان، در مقایسه با داده‌های نوشتاری، کاری بسیار دشوار است. زبان نوشتاری به راحتی قابل ذخیره در فایل‌های متنی الکترونیکی است. مشکل اساسی در مورد داده‌های گفتاری زبان، نمود نوشتاری متنی است که خوانده شده یا گفتاری است که در گذشته ضبط شده است. در حالیکه در مورد پیکره‌های متنی این مشکل وجود ندارد. زیرا کلمات موجود در یک پیکره متنی، قبل از ساخت پیکره وجود داشته‌اند. کلمات پیکره گفتاری تنها نمود بخشی از رخداد گفتار اصلی محسوب می‌شوند. تحلیلگر پیکره ناگزیر است گفتاری ضبط شده را آوانویسی کند و اطلاعات زبرنجیری به آن اضافه کند یا مشخصه‌های مربوط به بافت را هم در ضبط لحاظ کند. نه تنها آوانویسی گفتار، بلکه خود فرآیند جمع‌آوری داده‌های گفتاری، کاری بسیار دشوار و چالش‌برانگیز است. در ضبط صدا ویژگی‌های

بصری و حسی مغفول می‌ماند و امکان انعکاس آن‌ها در صدای ضبط شده وجود ندارد. البته می‌توان از ویدئو برای ضبط استفاده کرد، اما در هر حال نظر شرکت‌کنندگان در این امر، ضبط صدا/ تصویرشان و بیان حس آن‌ها مغفول می‌ماند. به طور کلی این نکته را باید در نظر داشت که ساخت پیکره گفتاری باید تحت شرایط خاص و با توجه به هدف ساخت پیکره صورت پذیرد. به عنوان مثال، چنانچه زبان‌شناسی قصد بررسی فراوانی واژگانی کلمات در حجم عظیمی از داده‌های زبانی را داشته باشد، نیازی به آوانویسی دقیق و موشکافانه گفتار نیست و مسئله اصلی کمیت و سرعت آوانویسی است و بنابراین آوانویسی ساده کفایت می‌کند؛ در حالیکه یک آواشناس برای بررسی آوایی داده‌های گفتاری، نیازمند داده‌های کمتری است، اما این داده‌ها باید با دقت بسیار بالا همراه با جزئیات آوایی دقیق، آوانویسی شده باشند و امکان دسترسی به صدای ضبط شده مربوط به هر آوانویسی هم وجود داشته باشد. یک تحلیلگر گفتمان برای تجزیه و تحلیل داده‌های گفتاری زبان، نیازمند اطلاعات دقیق درباره ویژگی‌های بافتی است که گفتار/ مکالمه در آن صورت پذیرفته است. بنابراین، هدف مطالعه بر چگونگی ساخت پیکره گفتاری تاثیرگذار است. لیچ و همکاران (۱۹۹۵)، نقل از تامسون (۲۰۰۴) برای ساخت پیکره گفتاری پنج مرحله در نظر می‌گیرند: ۱- ضبط، ۲- آوانویسی، ۳- نمایش / نمود^۱، ۴- کدگذاری (یا حاشیه‌نویسی)، ۵- به کار بردن. درباره مرحله اول چند مسئله باید در نظر گرفته شود: ویژگی‌های تکنیکی ضبط صوتی / تصویری، جمع‌آوری اطلاعات بافتی و رضایت شرکت‌کنندگان؛ در نتیجه این مرحله را می‌توان «جمع‌آوری داده‌ها» نامید. پس از جمع‌آوری داده‌های گفتاری زبان، فرآیند آوانویسی آغاز می‌شود. مرحله سوم، تبدیل آوانویسی به شکلی است که برای رایانه قابل خواندن باشد. پس از این مرحله، تحلیلگر می‌تواند، به تناسب هدف ساخت پیکره، اطلاعات اضافی از قبیل طبقه‌بندی کنش‌های گفتاری داده‌ها، مشخص کردن طبقه دستوری کلمات و... را اضافه کند؛ این مرحله «حاشیه‌نویسی» خوانده می‌شود. تامسون (۲۰۰۴) دو مرحله اخیر را یک مرحله در نظر می‌گیرد و عنوان «نمایش و حاشیه‌نویسی» را برای آن انتخاب می‌کند. وی مرحله نهایی را «دسترس^۲» می‌نامد؛ زیرا تاکید روی دسترسی به پیکره است و نه لزوماً به کار بردن آن. نمی‌توان دقیقاً رویکردهای مربوط به کار بردن چنین پیکره‌هایی را مورد بررسی قرار داد. اما می‌توان بررسی کرد آیا پیکره در دسترس دیگر پژوهشگران قرار می‌گیرد؟ و به چه شکلی در دسترس است؟ بنابراین، با در نظر گرفتن تغییرات مذکور، تامسون (۲۰۰۴)، چهار مرحله برای ساخت پیکره گفتاری

¹ representation

² access

تعریف می‌کند: ۱- جمع‌آوری داده‌ها، ۲- آوانویسی، ۳- نمایش و حاشیه‌نویسی، ۴- دسترسی. هر کدام از این مراحل توضیح داده می‌شوند:

۳-۴-۱- جمع‌آوری داده‌ها

قبل از جمع‌آوری داده‌ها، باید نحوه انجام این کار و مسائل مربوطه به شرکت‌کنندگانی که قرار است مکالمه آن‌ها با دیگران مورد استفاده قرار گیرد، توضیح داده شود و رضایت کتبی یا شفاهی مستند آن‌ها دریافت شود. این کار را می‌توان با ارائه فرم رضایت‌نامه پس از توضیحات مربوطه یا فیلم گرفتن از بیان شفاهی رضایت شرکت‌کنندگان، انجام شود. پس از این مرحله، ضبط گفتار آغاز می‌شود. پیشرفت تکنولوژی، انتخاب‌های گوناگونی برای ضبط گفتاری پیش روی پژوهشگران قرار داده است. طبق استانداردهای تعریف‌شده این حوزه، باید از هدفن، میکروفن، ویدئو، دوربین‌های کوچک و ابزارهای دیجیتال استفاده کرد. نکته مهم این است که ابزار مورد استفاده در این بخش باید به سهولت مورد استفاده قرار گیرند و شرکت‌کنندگان مشکلی در استفاده از این ابزارها نداشته باشند.

علاوه بر ضبط رخداد مکالمه / گفتار، باید اطلاعات پس‌زمینه و محیطی که مکالمه / گفتار در آن رخ می‌دهد نیز ثبت شود؛ مخصوصاً اگر قرار است فرآیند ضبط توسط افرادی انجام شود که لزوماً پژوهشگر نیستند. به عنوان مثال، پیکره ملی بریتانیایی، حاوی مکالماتی است که خود شرکت‌کنندگان با دوستان و همکاران خود داشته‌اند؛ این افراد اطلاعات گوینده و رضایت گویندگان را ضمیمه پروژه ساخت پیکره کرده‌اند (تامسون: ۲۰۰۴). نکته آخر اینکه، باید در ثبت جزئیات هر ضبط گفتار، دقت عمل کافی وجود داشته باشد تا همه فایل‌ها از نظر جزئیات اطلاعات در یک سطح و یک‌دست باشند.

۳-۴-۲- آوانویسی

ادوارد (۱۹۹۳، نقل از تامسون: ۲۰۰۴) سه اصل را برای آوانویسی دستی و کامپیوتری و طبقه‌بندی‌های مربوطه در نظر می‌گیرد: ۱- طبقه‌بندی‌ها باید مجزا از هم، قابل تشخیص و دقیق باشند، ۲- برای پژوهشگران قابل خواندن باشند و ۳- برای کامپیوتر باید سیستماتیک و قابل پیش‌بینی باشد. در آوانویسی نکته اساسی

این است که باید نوع آوانویسی مشخص باشد: وابسته به املای^۱ (در این حالت یک فرهنگ لغت را مرجع قرار داده می‌شود و از استانداردهای آوانویسی آن پیروی می‌شود و هر جا لازم باشد اطلاعات اضافی گنجانده می‌شود)، زبرنجیری (تکیه، آهنگ کلام و... نمایش داده شود) یا آوانگاری (استفاده از الفبای جهانی آوانگاری که زبان‌شناسان استفاده می‌کنند). می‌توان حتی بیش از یک نوع آوانویسی را در پیکره گنجانده، در این صورت باید انواع آوانویسی مربوط به هر کلمه در ستون‌ها یا ردیف‌های مجزا از هم قرار گیرند. در آوانویسی باید درباره چگونگی نمایش داده‌های غیرکلامی مانند اطلاعات بافتی، وقفه‌ها و مشخصه‌های فرازبانی نیز تصمیم‌گیری شود. به عنوان مثال، وقفه می‌تواند کوتاه، کمی طولانی یا زمان‌بر باشد. هر کدام را می‌توان به شکل‌های گوناگون در آوانویسی لحاظ کرد. جدول ۳-۶، مثالی را در این باره نشان می‌دهد:

جدول ۳-۶: چهار رویکرد نمایش نمادین وقفه و طول وقفه. (برگرفته از جانسون: ۱۹۹۵، نقل از تامسون: ۲۰۰۴)

Du Bois et al 1990	Short pause ..	Longer pause ...	Timed pause ... (1.5)
MacWhinney (1991)	Short pause #	Longer pauses ##, ###, #long	Timed pause #1_5
Rosta (1990)	Short pause <,>	Long pause <,,>	
Svartvik and Quirk (1980)	Brief pause .	Unit pause _	Longer pauses _· _·

۳-۴-۳- نمایش و حاشیه‌نویسی

باید روشی برای نشانه‌گذاری و حاشیه‌نویسی اتخاذ شود که مستقل از هر سیستم خاص کاربری باشد و در هر رایانه‌ای قابل خواندن باشد. زبان‌هایی مانند HTML و XML عموماً برای این منظور استفاده می‌شوند و کارایی بالایی دارند. باید از کاراکترهایی استفاده شود که خاص یک زبان خاص رایانه و سیستم خاص نباشد؛ معمولاً توصیه می‌شود از استانداردهای تعریف شده در TEI^۲ استفاده شود، این استانداردها

^۱ orthographic

^۲ Text Encoding Initiative

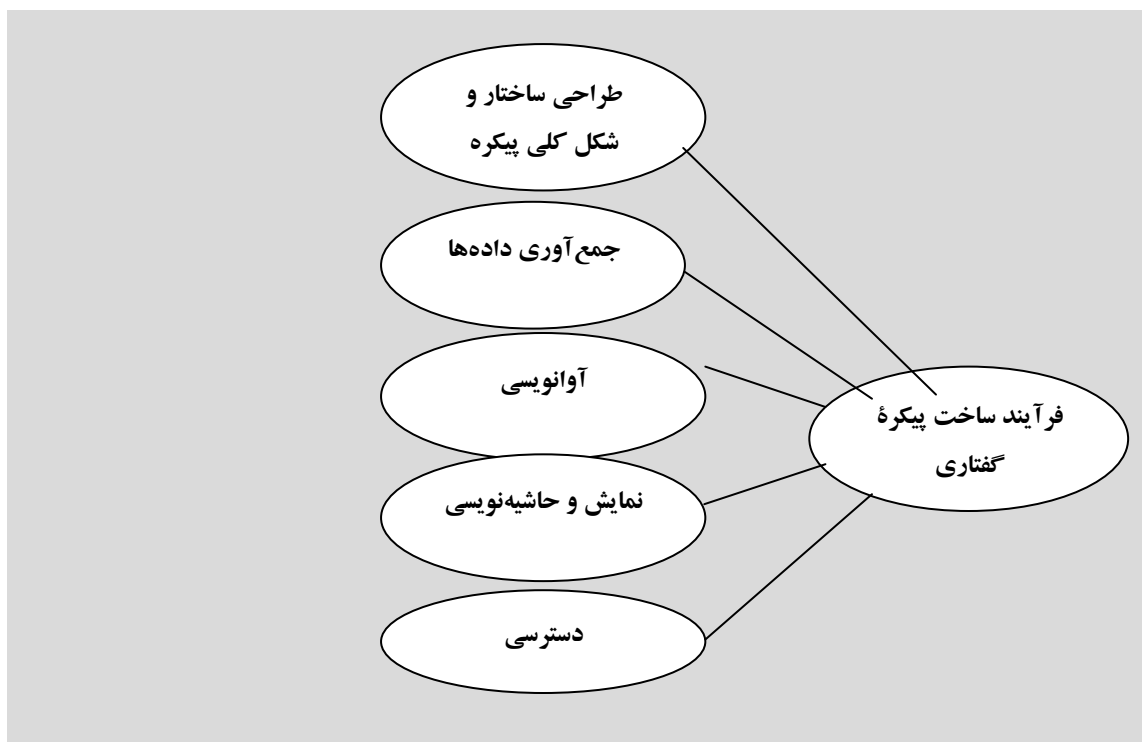
برای نشانه‌گذاری و حاشیه‌نویسی متون الکترونیکی تعریف شده است. مجموعه برچسب‌های TEI شامل: پاره‌گفتارها، وقفه‌ها، آواهای غیرواژگانی و زبانی مانند «سرفه»، رخدادهای غیرآوایی و غیرزبانی مانند «حرکات بدن»، موارد کاملاً غیرزبانی و محیطی، نوشتار، تغییر در کیفیت آوا است؛ این مجموعه برچسب‌ها، غیر از آواهای گفتاری است که ثبت و حاشیه‌نویسی می‌شوند. می‌توان موارد دیگر را نیز به آن‌ها اضافه کرد. البته می‌توان برخی دیگر از انواع حاشیه‌نویسی را به حاشیه‌نویسی مورد توافق سازندگان پیکره، تبدیل کرد. به طور کلی سؤالاتی که هنگام انتخاب حاشیه‌نویسی مطرح می‌شود، می‌تواند شامل این موارد باشد: آیا به راحتی می‌توان اطلاعات را از فایل‌های پیکره استخراج کرد؟ آیا برچسب‌ها یگانه و خاص هستند؟ آیا برچسب‌هایی وجود دارند که نیازمند اضافه کردن تفسیر برای هر نمونه/ مثال مربوطه باشند؟ آیا اینکه بتوان همه برچسب‌ها را به سرعت حذف کرد، کاری مفید تلقی می‌شود؟ آیا همه اطلاعات پس-زمینه (جزئیات مربوط به شرکت‌کنندگان، بافت و...) به راحتی در دسترس هستند؟

قبل از حاشیه‌نویسی داده‌های گفتاری زبان (چه دستی یا ماشینی)، در مقایسه با داده‌های نوشتاری، باید پیش‌پردازش‌های بیشتری روی داده‌ها صورت پذیرد. به عنوان مثال، یکی از پیش‌پردازش‌های خاص داده‌های گفتاری، تصمیم‌گیری درباره حذف یا تصحیح تکرار (مانند تکرار یک هجا یا یک کلمه در گفتار) یا عبارت‌های ناتمام است. نکته مهم این است که آوانویسی باید درست و طبق استانداردهای تعریف شده در زبان‌شناسی باشد تا به عنوان درونداد درست مورد استفاده قرار گیرد.

۳-۴-۴-دسترسی

آخرین مرحله از ساخت پیکره، فراهم کردن امکان دسترسی دیگران به پیکره است. می‌توان نسخه چاپی پیکره را در اختیار دیگران قرار داد، اما این امر تجزیه و تحلیل‌ها مبتنی بر پیکره را محدود می‌کند. بهتر است نسخه الکترونیکی منتشر شود تا به راحتی بتوان جستجو یا هر نوع تجزیه و تحلیل پیکره-بنیاد را انجام داد. در نسخه منتشر شده بهتر است غیر از آوانویسی گفتار، مستندات از قبیل: گفتار ضبط شده و اطلاعات مربوطه نیز وارد شود تا برای کاربر امکان مقایسه آوانویسی با فایل‌های صوتی/ تصویری فراهم شود؛ برای این منظور برنامه‌های رایانه‌ای/ نرم‌افزارهایی مانند Transana و Anvil وجود دارد که به تحلیلگر این امکان را می‌دهد فایل صوتی/ تصویری را به آوانویسی ارتباط دهد. نکات مطرح شده درباره ساخت پیکره گفتاری را می‌توان به صورت زیر خلاصه کرد (تامسون: ۲۰۰۴):

ساخت پیکره مبتنی بر داده‌های گفتاری، کاری بسیار پیچیده است که مستلزم برنامه‌ریزی دقیق است. باید این اطمینان حاصل شود که همه اطلاعات مربوطه جمع‌آوری شده است و از نظر کیفیت داده‌ها، فرآیند آوانویسی، پیروی از قراردادها و استانداردهای مربوطه، نمایش و حاشیه‌نویسی، یک‌دست هستند. باید درباره میزان داده‌های گفتاری جمع‌آوری شده و جزئیات حاشیه‌نویسی و میزان وارد کردن جزئیات، تصمیم‌گیری شود. در صورت وجود منابع کافی، بهتر است تا حد امکان داده‌ها بیشتری در ساخت پیکره وارد شوند و امکان نمایش آن‌ها به طرق گوناگون فراهم شود و حتی‌الامکان سعی شود بین فایل‌های صوتی/ تصویری و آوانویسی مربوطه، ارتباط ایجاد شود. فرآیند ساخت پیکره گفتاری را می‌توان به صورت شکل ۳-۳ نمایش داد:



شکل ۳-۳: فرآیند ساخت پیکره گفتاری

۳-۵- خلاصه فصل

در این فصل، ابتدا به شاخص‌های کلی ساخت پیکره‌های زبانی پرداخته شد. این شاخص‌ها شامل ساخت انواع پیکره متنی و گفتاری است که شامل: نمونه‌گیری، نمایندگی، توازن، اندازه، نوع پیکره و دسترسی است. سپس، فرآیند ساخت پیکره متنی ارائه شد که شامل: انتخاب متون، پیش‌پردازش متون و حاشیه‌نویسی است و در این راستا به تفصیل درباره هر یک از مراحل توضیح داده شده است. در پایان، فرآیند ساخت پیکره گفتاری بیان شده است که شامل: جمع‌آوری داده‌ها، آوانویسی، نمایش و حاشیه‌نویسی و دسترسی است. درباره هر یک از مراحل مذکور نیز به تفصیل توضیح داده شده است.

فصل چهارم

توسعهٔ چارچوب مفهومی امکان‌سنجی ساخت

پیکرهٔ زبانی

۴-۱- مقدمه

در این فصل ابتدا به تعریف امکان‌سنجی، سپس به بررسی ابعاد گوناگون امکان‌سنجی پرداخته خواهد شد و در نهایت مدل امکان‌سنجی ساخت پیکره زبانی ارائه خواهد شد.

۴-۲- تعریف امکان‌سنجی

امکان‌سنجی^۱ عبارت است از تجزیه و تحلیل قابلیت به عمل نشستن یک ایده در یک مسیر مشخص و منطقی و ترتیب و نظم دادن به مراحل انجام کار از ابتدا تا انتها (توماس و همکاران: ۱۹۸۲، نقل از شکری: ۱۳۹۴). در واقع انجام مطالعه امکان‌سنجی به معنی این است که مسئله اصلی، راه‌حل‌های مربوطه و نیازهای اولیه پروژه در نظر گرفته شده است و طی امکان‌سنجی به بررسی راه‌حل‌ها پرداخته می‌شود و در نهایت توصیه‌هایی مشخص ارائه می‌شود؛ این توصیه‌ها نشان می‌دهند آیا پروژه ادامه یابد یا متوقف شود؛ البته این توصیه‌ها تنها درباره امکان انجام یا ادامه پروژه نیست، بلکه درباره هزینه‌های مربوطه نیز اطلاعاتی در اختیار توسعه‌دهندگان پروژه قرار می‌دهد (پاور و همکاران: ۱۹۹۰). یکی از مهم‌ترین دستاوردهای مطالعات امکان‌سنجی دسترسی به اطلاعات حیاتی است که سازمان‌ها و مدیران پروژه را از خطر مردود شدن در نظر ذینفعانشان مصون نگه می‌دارد و پروژه را از شکست در اثر نادیده گرفتن چنین ریسک‌هایی دور نگه می‌دارد. پروژه‌هایی که در آن‌ها مطالعات امکان‌سنجی گسترده و باکیفیت صورت گرفته است، در فاز برنامه‌ریزی تفصیلی، مدیران پروژه با در نظر گرفتن اطلاعات به دست آمده در بخش امکان‌سنجی، می‌توانند تصمیمات درستی را برای تولید یک برنامه زمان‌بندی صحیح و با در نظر گرفتن ریسک‌های موجود اتخاذ نمایند. در امکان‌سنجی به بررسی امکان عملیاتی کردن یک طرح در یک زمان مشخص در یک سازمان پرداخته می‌شود. امکان‌سنجی یک روش مطالعه برای سنجش منافع، مزایا و هزینه‌های اجرای یک طرح است. هزینه و فایده، تنها معنا و دلالت مادی و مالی ندارد، بلکه در یک نگاه فراگیر و چندبعدی است که باید ابعاد مختلف و چندگانه یک طرح، اعم از مالی، انسانی، سازمانی، قانونی، امکانات نرم‌افزاری و سخت‌افزاری به صورت ترکیبی در نظر آورده شود. امکان‌سنجی در صدد شناخت وضع موجود و امکان اجرای طرح است. لذا تاثیرات ناشی از عملیاتی کردن و دگرگونی‌های احتمالی ناشی از اجرای یک ایده نوین را مورد سنجش قرار می‌دهد، اگر ایده‌ای برای اجرا مناسب

¹ feasibility study

تشخیص داده شود، مطالعه امکان‌سنجی، اطلاعات جدیدی را در مورد تامین مالی، جزئیات هزینه و غیره مهیا می‌سازد (سالاری: ۱۳۹۵). با مشخص شدن نتایج امکان‌سنجی، احتمال دارد از اجرای طرح صرف‌نظر گردد و گاه علیرغم اقتصادی بودن، ممکن است طرح به دلایل استراتژیک انتخاب و برای اجرا اقدام گردد. هر طرح باید برای اجرا و پیاده‌سازی از دیدگاه‌های مختلف بررسی شود؛ مانند اینکه آیا این طرح قابل اجرا است یا نه؟ چنانچه قابل اجرا نیست، دلایل آن چیست؟ در امکان‌سنجی طرح، کلیه توانایی‌ها و پتانسیل‌های سازمان مورد تجزیه و تحلیل قرار می‌گیرد (محمودی: ۱۳۹۰، نقل از سالاری: ۱۳۹۵).

۴-۳- بررسی پژوهش‌های امکان‌سنجی

تاکنون پژوهشی درباره امکان‌سنجی ساخت پیکره زبانی انجام نشده است، اما از پژوهش‌هایی که به طور کل، در حوزه امکان‌سنجی انجام شده است، می‌توان به پژوهش انجام شده توسط موخری و رئی (۲۰۱۷) اشاره کرد؛ آن‌ها ابتدا به بیان مزایای انجام مطالعات امکان‌سنجی می‌پردازند و پنج زمینه/بعد اصلی امکان‌سنجی پروژه را مورد بحث و بررسی قرار می‌دهند: امکان‌سنجی فنی، امکان‌سنجی اقتصادی، امکان‌سنجی قانونی/حقوقی، امکان‌سنجی عملیاتی و امکان‌سنجی زمان‌بندی. آن‌ها در نهایت عناصر یک امکان‌سنجی خوب و موفقیت‌آمیز را ذکر می‌کنند که خود شامل: اطمینان از صحت اطلاعاتی است که در امکان‌سنجی مورد استفاده قرار می‌گیرد، بیان درست و دقیق مسئله و در دست داشتن آخرین اطلاعات مستند از وضعیت مالی موجود برای انجام پروژه است. همچنین، در یک امکان‌سنجی خوب باید خطرات احتمالی قبل و بعد از انجام پروژه و احتمال برگرداندن سرمایه صرف شده را نیز در نظر گرفت. بوس و همکاران (۲۰۱۴) نیز مطالعات امکان‌سنجی را از پنج بعد مورد بررسی قرار می‌دهند: فنی، اقتصادی، قانونی/حقوقی، عملیاتی و زمان‌بندی. آن‌ها در این پژوهش بیشتر سعی دارند ارتباط میان امکان‌سنجی فنی و توسعه محصول را نشان دهند و در این راستا به جمع‌آوری جامع اهداف امکان‌سنجی فنی و ارائه مثال‌هایی از این نوع امکان‌سنجی در فرآیند توسعه محصول می‌پردازند و نتایج را در قالب اشکال گوناگون ارائه می‌دهند که امکان بررسی دقیق‌تر و مقایسه با تحقیقات مشابه را فراهم می‌آورد. یانگ و همکاران (۲۰۱۲) نتایج تجزیه و تحلیل حاصل از امکان‌سنجی اقتصادی یک مزرعه بادی را ارائه می‌دهند که در آن سه عامل هزینه تجزیه و تحلیل وضعیت موجود، یارانه اختصاص یافته از طرف دولت برای مزرعه‌های بادی و مکانیزم توسعه پاکیزه (clean development mechanism) مزرعه‌ها بادی در نظر گرفته می‌شود. بررسی نتایج نشان می‌دهد سیستم‌های تولیدکننده انرژی بادی، بهینه هستند و برای سرمایه‌گذاران جذاب به نظر می‌رسند.

جواهری (۱۳۸۳) با انجام امکان‌سنجی، سعی کرده است در جهت داشتن نگاهی واقع بینانه به محل، مشتریان بالقوه، منابع مالی و توسعه فعالیت‌های تفریحی گام بردارد. اهمیت و ضرورت مطالعات امکان‌سنجی، معرفی فعالیت‌های تفریحی و گردشگری، مراحل انجام مطالعات امکان‌سنجی و شیوه‌ها و شاخص‌های موجود مرتبط با این مطالعات و همچنین توجه به منابع و عوامل مکانی، سیمای مصرف‌کننده، سیمای ترکیبی مصرف‌کننده و ارزیابی فعالیت‌های تفریحی و گردشگری مورد نظر با تخمین هزینه‌ها، درآمدها، سودها و تدوین راهنمای اجرایی برای مطالعات امکان‌سنجی پروژه‌های تفریحی و گردشگری از جمله مواردی است که در این مقاله به آن‌ها اشاره شده است. ادیب‌منش و همکاران (۱۳۹۰) به امکان‌سنجی اجرای برنامه‌ریزی درسی مدرسه-محور، در دبیرستان‌های شهر کرمانشاه پرداخته‌اند. پژوهش آن‌ها توصیفی است و از پرسش‌نامه برای جمع‌آوری نظر صاحب‌نظران این حوزه استفاده کرده‌اند. در این امکان‌سنجی چهار متغیر دانش، نگرش، مهارت و امکانات در دو گروه مدیران و دبیران مورد سنجش و ارزیابی قرار داده‌اند. در واقع به نوعی، می‌توان گفت در این امکان‌سنجی ابعاد فنی و عملیاتی مورد بررسی قرار گرفته است. نتیجه کلی پژوهش حاکی از آن است که گروه‌های مخاطب دارای دانش کافی، نگرش متوسط و مهارت ناکافی در زمینه اجرای برنامه‌ریزی درسی مدرسه-محور می‌باشند و مدارس متوسطه شهر کرمانشاه، برای اجرای برنامه‌ریزی درسی مدرسه-محور، از امکانات مناسب برخوردار نیستند. مطالعه‌ای پورفرج و همکاران (۱۳۹۱) توصیفی-تحلیلی و از نوع کاربردی است که در آن به منظور سرمایه‌گذاری در شهرک تفریحی-توریستی بین‌المللی طارم (واقع در استان زنجان)، با استفاده از معیارهای مختلف ارزیابی مالی و اقتصادی و بهره‌گیری از نرم‌افزار comfar (مدلی رایانه‌ای برای تجزیه و تحلیل امکان‌سنجی و گزارش‌گیری)، به ارزیابی شاخص‌های اقتصادی پرداخته‌اند. نوع امکان‌سنجی آن‌ها در واقع، امکان-سنجی فنی، اقتصادی و بازاریابی است تا بوسیله آن بتوان درباره سرمایه‌گذاری روی چنین پروژه‌ای بهتر تصمیم‌گیری کرد. هدف از پژوهش نجفی‌خواه و همکاران (۱۳۹۵) امکان‌سنجی ارائه خدمات مرجع مجازی کتابخانه‌های عمومی استان گیلان وابسته به نهاد کتابخانه‌های عمومی کشور است. روش آن‌ها از نظر نوع، کاربردی و از نظر روش، پیمایشی است. مطابق مدل ارائه شده در این پژوهش، امکان‌سنجی فنی از چهار بعد سخت‌افزار، نرم‌افزار، شبکه و منابع و پایگاه‌های اطلاعاتی الکترونیکی بررسی شده است. همچنین در این پژوهش، وضعیت مالی موجود برای انجام پروژه نیز در نظر گرفته شده است و امکان-سنجی مالی از پنج بعد هزینه نرم‌افزار و پشتیبانی، هزینه نیروی انسانی، هزینه آموزش، هزینه سخت‌افزار،

هزینه منابع و پایگاه‌های اطلاعاتی الکترونیکی بررسی شده است. به نوعی امکان‌سنجی اقتصادی/ مالی و تامین منابع مالی نیز انجام شده است. ابزار گردآوری اطلاعات، دو پرسش‌نامه مبتنی بر پرسش‌نامه اربیس و ابزار ارزیابی انجمن کتابداران آمریکا است که یکی مخصوص سرپرستان کتابخانه‌ها و دیگری مخصوص کتابداران و مسئولان کتابخانه‌ها است. تجزیه و تحلیل داده‌ها با استفاده از شاخص‌های آمار توصیفی و نرم‌افزار اس. پی. اس. صورت گرفته است. نتایج پژوهش نشان می‌دهد که بیش از دو سوم کتابداران در جامعه آماری پژوهش، مهارت‌های لازم را در این زمینه دارند و در این زمینه امکان لازم فراهم است. اکثریت کتابخانه‌های مورد بررسی از نظر تجهیزات مورد نیاز برای راه‌اندازی خدمات مرجع مجازی با توجه به نرم‌افزار منتخب در پژوهش مشکلی ندارند. اما در زمینه منابع و پایگاه‌های اطلاعاتی الکترونیکی لازم برای تامین نیازهای اطلاعاتی کاربران، با مشکل مواجه هستند. فقدان بودجه مشکل اصلی کتابخانه‌ها در ارائه خدمات مرجع مجازی به شمار می‌رود. آن‌ها معتقدند کتابخانه‌های عمومی استان گیلان می‌توانند با تامین منابع و پایگاه‌های الکترونیکی لازم و نیز تامین بودجه مورد نیاز، اقدام به ارائه خدمات مرجع مجازی نمایند.

۴-۴-۱ ابعاد امکان‌سنجی

مطالعه امکان‌سنجی در سطوح مختلف انجام می‌گیرد؛ در واقع صرف آنکه مطالعه امکان‌سنجی انجام شود، نمی‌توان اطمینان لازم را برای یافتن بهترین پاسخ داشت. بنابراین، باید ابعاد و جنبه‌های مختلف موضوع مدنظر را تجسم و تبیین کرد. برخی از این ابعاد شامل موارد زیر است:

۴-۴-۱-۱ امکان‌سنجی فنی^۱

در این امکان‌سنجی، این مسئله بررسی می‌شود که آیا فناوری‌های لازم برای عملیاتی کردن طرح مورد نظر وجود دارد؟ منظور از فناوری، جنبه‌های سخت‌افزاری، نرم‌افزاری، انسانی و... است. به عبارت دیگر ارزیابی امکان‌سنجی فنی بر روی به دست آوردن یک درک از منابع فنی موجود در سازمان و امکان استفاده از آن‌ها در زمان نیاز در سیستم پیشنهادی تمرکز دارد و یک ارزیابی از سخت‌افزار و نرم‌افزار موجود و این که تا چه حد با سیستم پیشنهادی همخوانی دارند را فراهم می‌کند. به عنوان مثال، برای

¹ technical feasibility

امکان‌سنجی استفاده از سیستم‌های اطلاعاتی رایانه‌ای جدید در یک سازمان، بررسی می‌شود آیا تجهیزات و نرم‌افزارهایی که یک سازمان دارد یا می‌خواهد تهیه کند، توانایی لازم را برای پردازش اطلاعات و در نهایت کاربردهایی که برای آن‌ها تعریف شده است، دارد؟ (پاور و همکاران: ۱۹۹۰). در این نوع امکان-سنجی باید تکنولوژی مورد استفاده در سیستم مشخص گردد و در این راستا تکنولوژی‌های موجود در سازمان و نیازهای آموزشی نیز هم برای تولیدکنندگان نرم‌افزار و هم برای استفاده‌کنندگان می‌بایست مد نظر قرار گیرد. امکان‌سنجی فنی سه موضوع مهم و اصلی را مدنظر دارد (ویتنی: ۲۰۰۷، نقل از صرافی‌زاده و علیپور):

- آیا ما دارای تخصص فنی و دانش لازم جهت توسعه/ به کار بستن/ ساخت سیستم، هستیم؟
- آیا ما اکنون تکنولوژی لازم را داریم؟
- آیا تکنولوژی یا راه‌حل پیشنهادی، عملی است؟

مولفه‌های مطرح دیگر در این امکان‌سنجی شامل موارد زیر است (ppp-certification.com):

- آیا طراحی زیربنای پروژه، نیازهای مشخص‌شده سازمان را برآورده می‌سازد؟
- آیا پیش‌نیازهای معماری و مهندسی پروژه، دست‌یافتنی هستند؟
- آیا تکنولوژی پیشنهادی مورد سنجش قرار گرفته است؟ (البته این در صورتی است که تکنولوژی خاصی پیشنهاد شده باشد)
- ریسک‌های مربوطه قابل مدیریت هستند؟
- آیا می‌توان سرویس‌دهی را بر اساس خروجی مشخص کرد؟
- آیا می‌توان تغییرات اصلی فنی در زمینه ارائه سرویس را به نحو مطلوب تخمین زد؟

اگر پاسخ سوالات مذکور، مثبت باشد و هیچ مسئله فنی دیگری طی انجام امکان‌سنجی اضافه نشود، می‌توان گفت، پروژه از نظر فنی امکان‌پذیر است. کرنقان (۲۰۱۲) معتقد است تمرکز ارزیابی فنی یک امکان‌سنجی روی این مسئله است که آیا می‌توان یک پروژه را انجام داد؟ این به این معنی است که ارزیابی فنی یک امکان‌سنجی تنها برای پروژه‌های بسیار فنی و دشوار نیست، بلکه برای هر پروژه‌ای قابل اجرا است؛ این امکان‌سنجی به ارزیابی این مسئله نمی‌پردازد که آیا باید یک پروژه را انجام داد یا خیر؟

در ارزیابی فنی یک امکان‌سنجی، مولفه‌ها، معیارها و عوامل مختلف مورد بررسی قرار می‌گیرد؛ از میان آن‌ها می‌توان به موارد زیر، که طبقه‌بندی شده‌اند، اشاره کرد:

➤ تکنولوژی و تکنیک‌ها: سؤالاتی که در این حوزه از امکان‌سنجی فنی بررسی می‌شوند شامل این موارد است: آیا تکنولوژی و تکنیک‌های مورد نیاز پروژه در محل انجام آن یا به طور کل، موجود هستند؟ آیا قبلاً از این تکنولوژی و تکنیک‌های مربوطه استفاده شده است؟ چه کمک‌های خارج از سازمان برای استفاده از سیستم جدید یا پشتیبانی از استفاده از تکنولوژی و تکنیک‌های جدید، مورد نیاز است؟

➤ مهارت‌ها/ ظرفیت‌های فنی: سؤالاتی که در این حوزه از امکان‌سنجی فنی بررسی می‌شوند شامل این موارد است: آیا مهارت‌های لازم برای طراحی و اجرای پروژه داخل سازمان وجود دارد؟ اگر همه یا برخی از مهارت‌ها داخل سازمان وجود ندارد، چه مهارت‌هایی را باید بیرون از سازمان جستجو کرد و کجا می‌توان به آن‌ها دسترسی داشت؟

➤ محدودیت‌ها: چه محدودیت‌ها و ریسک‌هایی برای انجام پروژه وجود دارد؟

۴-۲-۱ امکان‌سنجی زمان‌بندی^۱

در صورتی که زمان تکمیل پروژه بیش از حد طولانی شود، قبل از اینکه مفید باشد ممکن است دچار شکست شود. این نوع امکان‌سنجی به این معنی است که تخمین بزنیم که توسعه یک سیستم چقدر طول می‌کشد؟ و آیا می‌توان آن را در یک بازه زمانی و با استفاده از برخی از روش‌ها، تکمیل نمود؟ آیا امکان تسریع در عملیات اجرایی وجود دارد؟ آیا امکان شناور ساختن پروژه وجود دارد؟ برای انجام یک پروژه، ممکن است تکنولوژی مربوطه را داشته باشیم، اما این به این معنی نیست که ما مهارت‌های لازم برای به کار بستن درست آن تکنولوژی را نیز داریم. درست است که همه متخصصان سیستم‌های اطلاعاتی قادرند تکنولوژی جدید را بیاموزند، اما این مسئله روی زمان‌بندی انجام پروژه تاثیر می‌گذارد. امکان‌سنجی زمان‌بندی مقیاسی است برای سنجش این مسئله که زمان‌بندی یک پروژه تا چه حد منطقی است؟ با داشتن توان فنی موجود، آیا موعدهای مقرر برای تحویل پروژه منطقی هستند؟ آیا موعد مقرر تحویل پروژه الزامی است یا مطلوب است؟ اگر مطلوب است که سر موعد مقرر تحویل داده شود، اما اجباری برای

¹ schedule feasibility

تحويل آن در آن زمان نیست، تحلیلگر می‌تواند زمان‌بندی جایگزینی را نیز پیشنهاد کند. اگر همان مثال استفاده از سیستم‌های اطلاعاتی رایانه‌ای جدید در یک سازمان را در نظر بگیریم، امکان‌سنجی زمان‌بندی زمانی انجام می‌شود که درخواست توسعه و طراحی یک سیستم جدید، مستلزم تعیین ضرب‌العجل خاص و غیرقابل تغییر باشد (پاور و همکاران: ۱۹۹۰). بهتر است سیستم اطلاعاتی، دو ماه دیرتر از موعد مقرر تحويل داده شود تا سیستم از نظر خطا نیز چک شده باشد. تحويل با تاخیر یک سیستم بی‌نقص یا کم‌خطا بهتر از تحويل به موقع یک سیستم حاوی خطاهای غیرقابل چشم‌پوشی است.

۴-۳-۴-۱ امکان‌سنجی اقتصادی^۱

هدف بررسی امکان‌سنجی اقتصادی، تعیین مزایای مثبت اقتصادی برای سازمانی است که سیستم پیشنهادی برای آن فراهم می‌شود و شامل اندازه‌گیری و شناسایی تمام مزایای مورد انتظار می‌شود. این ارزیابی معمولاً دارای یک تحلیل هزینه/فایده^۲ می‌شود. در واقع، هدف اصلی، مشخص کردن میزان صرفه‌جویی و مزایای دیگری است که در نتیجه به کار بستن یک سیستم جدید فراهم می‌شود. این مزایا با هزینه‌ها مقایسه می‌شود، اگر مزایا فراتر از حد کفایت باشد، تصمیم بر استفاده از سیستم پیشنهادی گرفته می‌شود، در غیر اینصورت، باید دلایل توجیه‌کننده دیگری برای توسعه سیستمی جدید وجود داشته باشد (پاور و همکاران: ۱۹۹۰). این امکان‌سنجی مهم‌ترین قسمت در مجموعه بررسی‌های امکان‌سنجی و بیشترین روش به کار رفته جهت ارزیابی تأثیر سیستم جدید است. در بسیاری از موارد، مسائل مالی مانع اصلی در انجام طرح‌ها بوده و حتی مجالی برای بررسی‌های فنی و عملیاتی باقی نمی‌گذارد. معمول‌ترین تجزیه و تحلیل شناخته شده در این حوزه، روش تجزیه و تحلیل هزینه/فایده است و روش کار بدین صورت است که در آن سود و صرفه‌جویی‌هایی که از یک سیستم معین انتظار می‌رود را مشخص می‌کنند و آن را با هزینه‌ها مقایسه می‌کنند. اگر سود بر هزینه‌ها چیرگی داشت، آن وقت تصمیم به طراحی و پیاده‌سازی سیستم گرفته می‌شود. یک مؤسس پیش از اقدام به هر عملی باید به درستی هزینه را در مقابل سود بسنجد. در واقع در این نوع امکان‌سنجی تلاش می‌شود تا موضوعات مهمی مانند فراهم بودن امکانات مالی لازم و چگونگی دسترسی به وجوه نقدی کافی به هنگام پیاده کردن طرح، مورد بررسی قرار گیرد

^۱ economic feasibility

^۲ cost-benefit

(شرف: ۱۳۸۶، نقل از سالاری: ۱۳۹۵). به عنوان مثال، برخی از مولفه‌هایی که در امکان‌سنجی اقتصادی سیستم‌های اطلاعاتی لحاظ می‌شود شامل موارد زیر است (همتی و همکاران: ۱۳۹۳):

- آیا مزایای حاصل از پیاده‌سازی سیستم اطلاعاتی از هزینه‌های اجرایی آن بیشتر است یا خیر؟
- آیا از نظر هزینه، سیستم اطلاعاتی اثربخش است؟
- آیا منابع مالی مورد نیاز برای پیاده‌سازی سیستم، در دسترس است؟
- هزینه نرم‌افزارها و سخت‌افزارهای مورد نیاز چقدر است؟
- هزینه نسبی مشاوره طرح / پروژه چقدر است؟
- هزینه‌های پنهان مانند وقت مجریان، مطالعه مقدماتی و آموزش تا چقدر است؟

برخی از مولفه‌های دیگر امکان‌سنجی اقتصادی شامل موارد زیر است:

- هزینه پرسنل / تیم اجرایی
- هزینه آموزش (چنانچه به کار بستن سیستم، مستلزم آموزش باشد)
- هزینه تجهیزات (هزینه نرم‌افزارها و سخت‌افزارهای مورد استفاده در پروژه)
- هزینه اسباب و اثاثیه جدید برای مکان انجام پروژه (رایانه، میز، صندلی و...)

۴-۴-۴- امکان‌سنجی عملیاتی^۱

امکان‌سنجی عملیاتی، مقیاسی است از اینکه یک سیستم پیشنهادی چطور مسائل را حل می‌کند، از فرصت‌های شناسایی شده در تعریف حوزه مورد بررسی، بهره می‌برد و چقدر الزامات شناسایی شده در فاز تحلیل، الزامات توسعه سیستم را برآورده می‌کند. ارزیابی امکان‌سنجی عملیاتی بر میزان مناسب بودن پروژه‌های توسعه‌ای پیشنهادی برای محیط کسب و کار موجود و اهداف، با توجه به توسعه زمان‌بندی، تاریخ تحویل، فرهنگ همکاری و فرآیندهای موجود کسب و کار تمرکز دارد. این امکان‌سنجی، درجه اهمیت مسئله و لزوم حل آن را اندازه‌گیری می‌کند. در واقع به بررسی طرح و این امر می‌پردازد که آیا طرح و ایده مورد نظر را می‌توان در بستر سازمان، تشکیلات و محیط موجود پیاده کرد و آیا با آن همخوانی سازمانی، فرهنگی و غیره دارد یا خیر؟ (شرف: ۱۳۸۶، نقل از سالاری: ۱۳۹۵). امکان‌سنجی عملیاتی باید نیازهای تجارت را برآورده سازد یا مشکلات آن را حل کند. به عنوان مثال، یکی از مسائلی

¹ operational feasibility

که در ارزیابی امکان‌سنجی سیستم پیشنهادی لحاظ می‌شود، این است که آیا سازمان می‌تواند پردازش خودکار اطلاعات را با کارایی بالا انجام دهد؟ آیا سیستم پیشنهادی می‌تواند، به جای پردازش رایانه‌ای، از نیروی انسانی برای پردازش استفاده کند؟ چطور سیستم پیشنهادی، کارگران، مدیران آن‌ها و سازمان را تحت تاثیر قرار می‌دهد؟ (پاور و همکاران: ۱۹۹۰). شید (۲۰۱۳) معتقد است پس از انجام امکان‌سنجی زمان‌بندی، اقتصادی و فنی، می‌توان امکان‌سنجی عملیاتی انجام داد تا مشخص شود آیا طراحی محیط کار مفید واقع خواهد شد یا خیر؟ وی معتقد است این امکان‌سنجی سعی دارد به سوالات زیر با تفکیک حوزه‌های مربوطه پاسخ دهد:

- حوزه فرآیند: احساس کاربران نهایی درباره انجام یک فرآیند جدید چه خواهد بود؟
- حوزه ارزیابی: آیا فرآیند درون سازمان عمل خواهد کرد/ می‌تواند عمل کند؟
- حوزه به کار بستن/ اجرا: وظایف مدیران و کاربران نهایی چه خواهد بود؟
- حوزه مقاومت: پس از ارزیابی مقاومت افراد، تیم و مدیریت، باید به این سوال پاسخ داد که چگونه می‌توان مقاومت آن‌ها را مورد رسیدگی قرار داد؟
- حوزه استراتژی‌های درون‌سازمانی: چطور محیط کاری تحت تاثیر قرار خواهد گرفت؟ تا چه اندازه تغییر خواهد کرد؟
- آیا سیستم پیشنهادی به نفع سازمان است؟
- آیا سیستم روی کاربران/ مشتریان تاثیر قابل ملاحظه خواهد داشت؟

۴-۵-۴- امکان‌سنجی عوامل انسانی^۱

اصولاً مردم اغلب نسبت به تغییرات، مقاوم نشان می‌دهند. همیشه احتمال وقوع نزاع بر سر توسعه و به کار بستن سیستم‌های جدید وجود دارد. به همین دلیل، امکان‌سنجی باید کاهش احتمال تاثیر عوامل انسانی را مورد ارزیابی قرار دهد. باید میزان مقاومت و مخالفت کارکنان با سیستم جدید تخمین زده شود و بررسی شود آیا عکس‌العمل مردم نسبت به استفاده از یک سیستم جدید، باعث به تاخیر انداختن یا مانع ایجاد کردن بر سر راه توسعه سیستم پیشنهادی می‌شود؟ اگر اینگونه باشد باید به ارزیابی حد و اندازه موانع پرداخته شود تا بتوان آن‌ها را برطرف ساخت (پاور و همکاران: ۱۹۹۰). با استفاده از روش‌هایی مانند

¹ human factors feasibility

مصاحبه، مشاهده و تنظیم پرسش‌نامه، می‌توان از کم و کیف مسائل انسانی و عواملی که بر اثر رایانه‌ای شدن نوع کار، تغییرات بسیاری ایجاد می‌کنند، اطلاع کسب کرد (فرشیدنژاد: ۱۳۷۹، نقل از سالاری: ۱۳۹۵).

۴-۴-۶- امکان‌سنجی قانونی/حقوقی^۱

امکان‌سنجی قانونی شامل اندازه‌گیری چگونگی اجرای بهینه یک راه‌حل درون الزامات قانونی موجود است (ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور). این نوع امکان‌سنجی تعیین می‌کند که آیا سیستم پیشنهادی با الزامات قانونی تناقضی دارد؟ به عنوان مثال، یک سیستم پردازش داده باید با قوانین حفاظت اطلاعات محلی همخوانی داشته باشد. این مرحله شامل بررسی محدودیت‌ها و موانع قانونی برای پیاده‌سازی سیستم می‌باشد و در این مرحله مواردی نظیر وجود کپی‌رایت‌ها یا جلوگیری از به کار بردن جملات یا کلمات متناقض و مبهم در قرارداد، لحاظ می‌شود. معمولاً این مرحله با مشاوره کارشناسان حقوقی انجام می‌شود و برای همه سیستم‌ها، خصوصاً سیستم‌های کوچک و ساده، مورد نیاز نیست. هدف اصلی امکان‌سنجی قانونی اطمینان از عدم ممانعت قانونی برای اجرای پروژه و اجتناب از مشکلات اساسی در توسعه و انجام پروژه است. مسائلی که در این امکان‌سنجی مورد بررسی قرار می‌گیرند شامل موارد زیر است (ppp-certification.com):

- در نظر گرفتن قوانین وضع شده مربوط به سرمایه‌گذاری خارجی و مسائل کاری
- در نظر گرفتن قوانین وضع شده درباره برنامه‌ریزی استفاده از زمین و قوانین مربوط به محیط زیست
- در نظر گرفتن جوانب قانونی نحوه تفکیک مباحث
- در نظر گرفتن قوانین مربوط به اعطای حق مالکیت یا کنترل دارایی عمومی یا قوانین مربوط به مسوولیت تحویل سرویس‌های عمومی به شخص ثالث
- در نظر گرفتن چگونگی برخورد قانونی با منابع درآمد مربوط به اعطای زمین و....

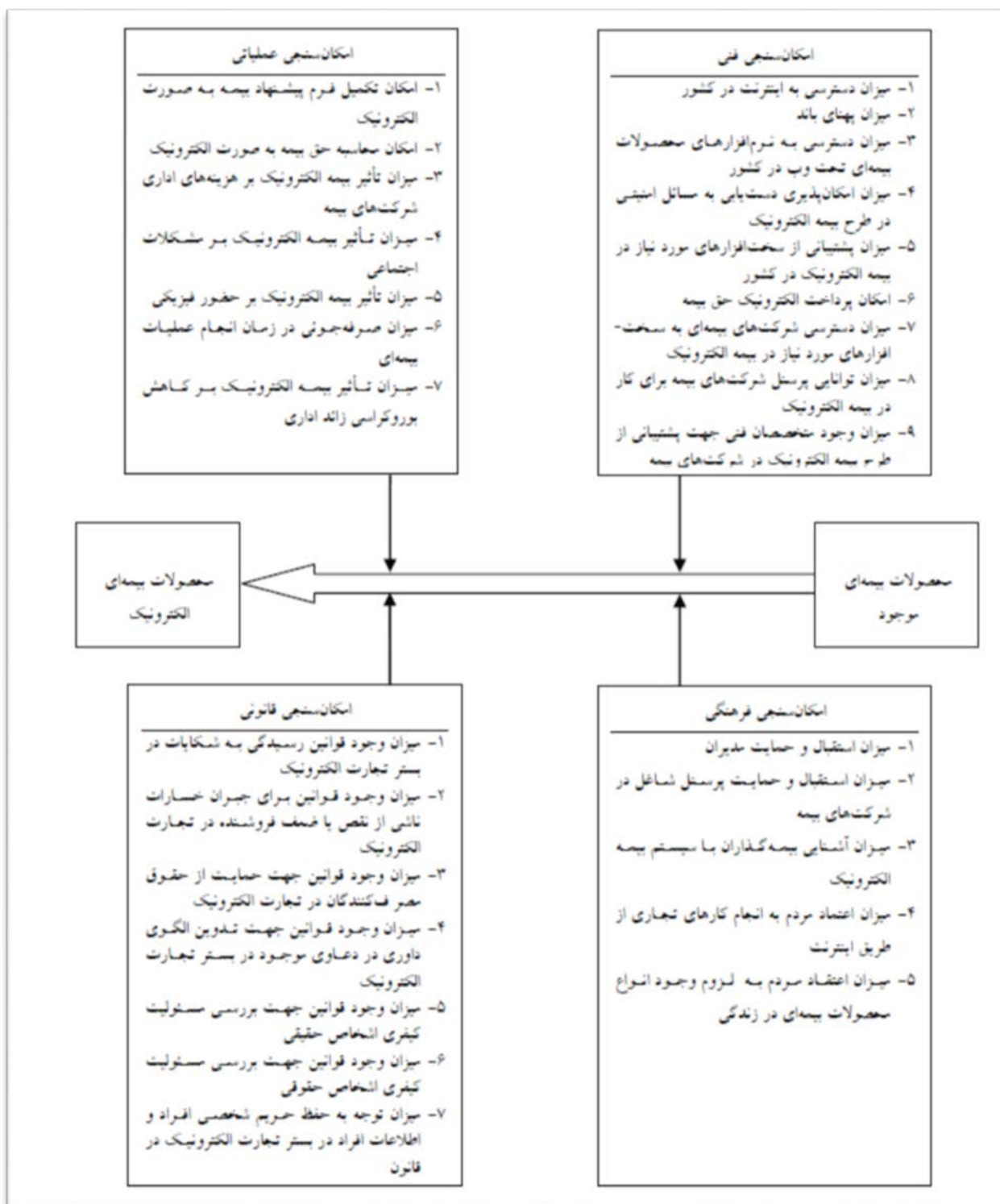
¹ legal feasibility

۴-۴-۷-۱ امکان‌سنجی فرهنگی^۱

این امکان‌سنجی، به چگونگی احساس کاربران نهایی در خصوص سیستم پیشنهادی، می‌پردازد. البته این مولفه در امکان‌سنجی عملیاتی نیز مورد بررسی قرار می‌گیرد. می‌توان گفت امکان‌سنجی فرهنگی به بررسی این مسئله می‌پردازد که آیا سیستم پیشنهادی در جو سازمانی مورد نظر مورد پذیرش واقع خواهد شد یا خیر؟ (ویتنی: ۲۰۰۷، نقل از صرافی‌زاده و علیپور: ۱۳۹۰).

به عنوان مثال، مدل امکان‌سنجی صرافی‌زاده و علیپور (۱۳۹۰) که برای ارائه محصولات بیمه‌ای موجود به صورت الکترونیک در برخی از شرکت‌های بیمه، طراحی شده است، شامل مولفه‌ها و ابعاد زیر است:

¹ cultural feasibility



شکل ۴-۱: مدل امکان‌سنجی ارائه محصولات بیمه‌ای موجود به صورت الکترونیک در برخی از شرکت‌های بیمه (صرافی -

زاده و علیپور: ۱۳۹۰)

موارد امکان‌سنجی ذکر شده معمول‌ترین ابعاد امکان‌سنجی هستند، البته موارد دیگر از انواع امکان‌سنجی نیز وجود دارد مانند موارد زیر:

- امکان‌سنجی منابع^۱: چه منابعی برا انجام پروژه لازم است؟ آیا منابع کافی داریم؟ چه امکاناتی برای انجام پروژه لازم است؟ (البته ممکن است این امکان‌سنجی با مولفه‌های امکان‌سنجی فنی همپوشانی داشته باشد). ppp.evirtualservices.com/resource-feasibility
- امکان‌سنجی بازاریابی^۲: آیا کسی خواهان محصول، پس از ساخت آن است؟ آیا محصول باید ابتدا مورد آزمایش قرار گیرد و سپس به بازار عرضه شود؟ ppp.evirtualservices.com/market-feasibility
- امکان‌سنجی جامع^۳: این امکان‌سنجی مستلزم بررسی جوانب مختلف مربوط به پروژه است که شامل بازاریابی، فرهنگی، اقتصادی و.... است. این امکان‌سنجی معمولاً هنگام راه‌اندازی کسب و کار جدید اعمال می‌شود.

معروف‌ترین ابعاد امکان‌سنجی‌ها در شکل ۴-۲ آورده شده است.

¹ resorce feasibility study

² marketing feasibility study

³ comprehensive feasibility study

امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۷۴



شکل ۴-۲: ابعاد امکان‌سنجی

۴-۵- اعتباریابی چارچوب مفهومی امکان‌سنجی ساخت پیکره

بر اساس دانش قبلی و بررسی متون جامعی که در پایگاه داده‌های داخلی و خارجی صورت گرفت، مطالعه‌ای جهت امکان‌سنجی ساخت پیکره زبانی انجام نشده است و پژوهش حاضر اولین مطالعه‌ای است که به بررسی این مسئله پرداخته است. به منظور طراحی مدل امکان‌سنجی ساخت پیکره زبانی و انتخاب ابعاد و مولفه‌های مناسب برای این نوع امکان‌سنجی، باید از نظر متخصصان ساخت پیکره استفاده کرد؛ در این راستا می‌توان از روش دلفی^۱ استفاده کرد. کاربرد روش دلفی، ساخت دادن به فرآیند ارتباطات میان افراد است به نحوی که چنین فرآیندی در فراهم کردن زمینه درگیری آنان به عنوان یک کل، با مسئله یا موضوعی پیچیده مؤثر باشد. این ارتباط ساخت یافته با بازخوران اطلاعات و دانش افراد، ارزیابی قضاوت یا نظر گروه، ارائه فرصتی به افراد برای بازنگری در نظرشان، و تا حدی پنهان نگه داشتن پاسخ‌های افراد از یک‌دیگر تأمین می‌شود (لینستون و تروف: ۱۹۷۵، نقل از علیدوستی: ۱۳۸۵). روش دلفی در عمل، مجموعه‌ای از پرسش‌نامه‌ها یا دورهای^۲ پشت سر هم به همراه بازخوران کنترل شده‌ای است که تلاش دارد به اتفاق نظر میان یک گروه از افراد خبره^۳ درباره یک موضوع خاص دست یابد. روش دلفی علاوه بر پیش‌بینی^۴، در گردآوری داده‌های مربوط به زمان حال یا گذشته که به درستی معلوم یا موجود نیستند و یافتن روابط علی در پدیده‌های پیچیده اجتماعی و اقتصادی نیز استفاده می‌شود (لینستون و تروف: ۱۹۷۵، و مک‌کنا: ۲۰۰۰، نقل از علیدوستی: ۱۳۸۵). این روش را می‌توان در مراحل گوناگون فرآیند یک پژوهش نیز مانند دستیابی به یک دیدگاه نظری برای پژوهش، گزینش متغیرها، تخمین مقدماتی روابط علی میان متغیرها، و تعریف سازه‌ها به کار برد. در هر صورت برخی از شرایطی که از روش دلفی استفاده می‌شود شامل موارد زیر است (لینستون و تروف: ۱۹۷۵، نقل از علیدوستی: ۱۳۸۵):

- مسئله از طریق فنون تحلیلی دقیق قابل بررسی نباشد، اما بتوان برای آن از قضاوت‌های انتزاعی^۵ یک گروه بهره برد.
- افرادی که باید در بررسی یک مسئله گسترده یا پیچیده مشارکت کنند، ارتباط کافی نداشته و احتمالاً دارای زمینه‌های تخصصی و تجربی متنوعی باشند.

^۱ Delphi method

^۲ rounds

^۳ expert panel

^۴ forecasting

^۵ subjective judgments

- تعداد افراد مورد نیاز بیش از آن باشد که بتوانند برهم کنش رودررو داشته باشند.
- برگزاری نشست‌های گروهی متناوب، از لحاظ زمان و هزینه توجیهی نداشته باشد.
- برای اطمینان از اعتبار نتایج، لازم باشد که عدم تجانس افراد حفظ گردد.

از سوی دیگر باید توجه داشت که این روش در صورتی قابل انجام است که متخصصان واجد شرایط به تعداد کافی برای انجام پژوهش وجود داشته و این افراد نسبت به شرکت در فرآیند دلفی تا دستیابی به اتفاق نظر متعهد باشند (هاسون و همکاران: ۲۰۰۰، نقل از علیدوستی: ۱۳۸۵).

به طور کلی روش دلفی شامل چند مرحله اساسی است (میتچل: ۱۹۹۱، جونز و هانتز: ۱۹۹۵، کلایتون: ۱۹۹۷، هاسون و همکاران: ۲۰۰۰، اکولیا و پالوسکی: ۲۰۰۴، نقل از علیدوستی: ۱۳۸۵): در مرحله اول، مسئله پژوهش تعریف و بر این اساس ویژگی‌های لازم برای شرکت‌کنندگان در پنل دلفی تعیین می‌شوند. سپس نامزدهای مشارکت در این پنل شناسایی و از آنان دعوت به عمل می‌آید. این مرحله با تعیین اعضای پنل به اتمام می‌رسد. مرحله دوم روش دلفی به تولید ایده در زمینه مسئله پژوهش اختصاص دارد. در این مرحله، اعضای پنل ایده‌های خود را درباره عوامل مرتبط با مسئله پژوهش ارائه می‌کنند. پژوهشگر با تحلیل و پالایش این ایده‌ها، حذف موارد تکراری، و کاربرد واژگان یکسان، فهرست نهایی عوامل مرتبط با مسئله پژوهش را استخراج می‌کند. در این مرحله ممکن است پژوهشگر، خود عواملی را از پیش تعیین و آن‌ها را برای نظرخواهی در اختیار اعضا قرار دهد. در مرحله دوم می‌توان از هر دو روش هم استفاده نمود، به این معنا که علاوه بر دریافت ایده‌های اعضای پنل، نظر آن‌ها را نیز درباره عواملی که از پیش تعیین شده‌اند، درخواست کرد. فرآیند تولید ایده‌های جدید از دور اول روش دلفی آغاز می‌شود. در این دور از اعضای پنل خواسته می‌شود که آزادانه نظر خود را در پاسخ به پرسش‌های مطرح شده، ارائه کنند. هر چند از لحاظ نظری نباید محدودیتی برای تعداد ایده‌ها قائل شد تا امکان دستیابی به مهم‌ترین ایده‌ها یا موضوع‌ها بیشینه شود، اما در عمل تعداد زیاد ایده‌ها ممکن است دشواری‌هایی ایجاد کند. از این رو برخی از پژوهشگران دست کم شش ایده جدید را کافی می‌دانند. علاوه بر این در دور اول باید دقت کافی برای تدوین پرسش‌نامه صرف شود تا از درستی نگارش و جمله‌بندی، و قابلیت درک پرسش‌ها اطمینان حاصل گردد. تحلیل داده‌های پرسش‌نامه‌ها دربردارنده تحلیل و مدیریت دقیق داده‌های کمی و کیفی است. داده‌های حاصل از دور اول دلفی اغلب کیفی هستند و می‌توان آن‌ها را با فنونی مانند تحلیل محتوا، تحلیل کرد. داده‌های این دور با گروه‌بندی اقلام مشابه در کنار هم تحلیل می‌شوند. در این مرحله واژه‌های گوناگونی که برای یک مقوله واحد به کار رفته‌اند، از سوی پژوهشگر برای ارائه یک توصیف

واحد، در یک گروه قرار می‌گیرند. چنین توصیف‌ها و گروه‌بندی‌ها باید به گونه‌ای صورت پذیرند که از جمع‌بندی و ارائه صحیح داده‌ها اطمینان وجود داشته شود. در مرحله سوم، اعضای پنل میزان اهمیت عوامل را تعیین یا تعدادی از مهمترین آن‌ها را انتخاب می‌کنند. بر این اساس، تعداد عوامل به میزانی کاهش می‌یابند که کار با آن‌ها قابل انجام باشد. در حقیقت، این مرحله برای کاهش تعداد عوامل به تعداد قابل قبول برای ادامه کار انجام می‌شود. مرحله چهارم به بازنگری در میزان اهمیت عوامل بر اساس نتایج مرحله پیش یا تعیین ترتیب اهمیت عوامل اختصاص دارد. در این مرحله، هر یک از اعضا در جریان نظر گروه و نظر پیشین خود قرار می‌گیرد و در میزان یا ترتیب اهمیت عوامل تجدید نظر می‌کند. تجدید نظر اعضا تا جایی ادامه می‌یابد که میان آنان اتفاق نظر حاصل شود یا تعداد شرکت‌کنندگان به کمتر از حد لازم برسد.

۴-۵-۱- فرآیند ساخت پرسش‌نامه

در پژوهش حاضر، برای طراحی مدل امکان‌سنجی ساخت پیکره با استفاده از روش دلفی، به این صورت عمل شد که ابتدا اطلاعات بخش انواع امکان‌سنجی و ابعاد مربوطه، به صورت زیر جمع‌بندی شد:

• جمع‌بندی محتوای ابعاد:

➤ بعد فنی: در این سازه از امکان‌سنجی، این مسئله بررسی می‌شود که آیا فناوری‌های لازم برای عملیاتی کردن طرح مورد نظر وجود دارد؟ منظور از فناوری، جنبه‌های سخت‌افزاری، نرم‌افزاری، انسانی و... است. به عبارت دیگر ارزیابی امکان‌سنجی فنی بر روی به دست آوردن یک درک از منابع فنی موجود در سازمان و امکان استفاده از آن‌ها در زمان نیاز در سیستم پیشنهادی تمرکز دارد و یک ارزیابی از سخت‌افزار و نرم‌افزار موجود و این که تا چه حد با سیستم پیشنهادی همخوانی دارند را فراهم می‌کند (پاور و همکاران: ۱۹۹۰).

➤ بعد اقتصادی: هدف بررسی اقتصادی، تعیین مزایای مثبت اقتصادی برای سازمانی است که سیستم پیشنهادی برای آن فراهم می‌شود و شامل اندازه‌گیری و شناسایی تمام مزایای مورد انتظار می‌شود. در این امکان‌سنجی تلاش می‌شود تا موضوعات مهمی مانند فراهم بودن امکانات مالی لازم و چگونگی دسترسی به وجوه نقدی کافی به هنگام پیاده کردن طرح، مورد بررسی قرار گیرد (پاور و همکاران: ۱۹۹۰) و (شرف: ۱۳۸۶، نقل از سالاری: ۱۳۹۵).

- بعد زمان‌بندی: مقیاسی است برای سنجش این مسئله که زمان‌بندی یک پروژه تا چه حد منطقی است و با داشتن توان فنی موجود، آیا موعدهای مقرر برای تحویل پروژه منطقی هستند؟ (پاور و همکاران: ۱۹۹۰).
- بعد عملیاتی: مقیاسی است از اینکه یک سیستم پیشنهادی چطور مسائل را حل می‌کند، از فرصت‌های شناسایی شده در تعریف حوزه مورد بررسی، بهره می‌برد و چقدر الزامات شناسایی شده در فاز تحلیل، الزامات توسعه سیستم را برآورده می‌کند. ارزیابی امکان‌سنجی عملیاتی بر میزان مناسب بودن پروژه‌های توسعه‌ای پیشنهادی برای محیط کسب و کار موجود و اهداف، با توجه به توسعه زمان‌بندی، تاریخ تحویل، فرهنگ همکاری و فرآیندهای موجود کسب و کار تمرکز دارد. این امکان‌سنجی، درجه اهمیت مسئله و لزوم حل آن را اندازه‌گیری می‌کند (شرف: ۱۳۸۶، نقل از سالاری: ۱۳۹۵).
- بعد عوامل انسانی: در این سازه، کاهش احتمال تاثیر عوامل انسانی مورد ارزیابی قرار می‌گیرد. میزان مقاومت و مخالفت کارکنان با سیستم جدید تخمین زده می‌شود (پاور و همکاران: ۱۹۹۰).
- بعد قانونی / حقوقی: این سازه شامل اندازه‌گیری چگونگی اجرای بهینه یک راه‌حل درون الزامات قانونی موجود است. هدف اصلی امکان‌سنجی قانونی اطمینان از عدم ممانعت قانونی برای اجرای پروژه و اجتناب از مشکلات اساسی در توسعه و انجام پروژه است (ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور) و (ppp-certification.com).
- بعد فرهنگی: این سازه از امکان‌سنجی، به چگونگی احساس کاربران نهایی در خصوص سیستم پیشنهادی، می‌پردازد. می‌توان گفت امکان‌سنجی فرهنگی به بررسی این مسئله می‌پردازد که آیا سیستم پیشنهادی در جو سازمانی مورد نظر مورد پذیرش واقع خواهد شد یا خیر؟ (ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور: ۱۳۹۰).
- بعد منابع: به چگونگی احساس کاربران نهایی در خصوص سیستم پیشنهادی، می‌پردازد. می‌توان گفت امکان‌سنجی فرهنگی به بررسی این مسئله می‌پردازد که آیا سیستم پیشنهادی در جو سازمانی مورد نظر مورد پذیرش واقع خواهد شد یا خیر؟ (ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور: ۱۳۹۰).
- بعد بازاریابی: این امکان‌سنجی به بررسی بازار احتمالی محصول پس از ساخت آن می‌پردازد.

ابعاد و مؤلفه‌هایی را که در پژوهش‌های پیشین دربارهٔ امکان‌سنجی پیدا شده‌اند، می‌توان در جدول ۴-۱ دید.

جدول ۴-۱، ابعاد امکان‌سنجی و مؤلفه‌های مربوط به هر یک از ابعاد

ابعاد	معیارها/ مقیاس‌ها	منابع
۱. فنی	۱-۱. مهارت‌ها و تخصص‌های فنی و دانش لازم برای ساخت/ توسعه/ کاربرست سیستم ۲-۱. فناوری و فنون لازم برای ساخت/ توسعه/ کاربرست سیستم ۳-۱. کمک‌های بیرون از سازمان برای کاربرد سیستم تازه یا پشتیبانی از فناوری و فنون ۴-۱. طراحی زیرساخت پروژه ۵-۱. پیش‌نیازهای معماری و مهندسی پروژه ۶-۱. سخت‌افزار، نرم‌افزار، شبکه، و منابع و پایگاه‌های اطلاعاتی الکترونیکی	(ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور: ۱۳۹۰)، «موخری» و «رئی» (۲۰۱۷)، «بوس» و همکاران (۲۰۱۴)، «یانگ» و همکاران (۲۰۱۲)، کرنقان (۲۰۱۲)، «ادیب‌منش» و همکاران (۱۳۹۰)، «پورفرج» و همکاران (۱۳۹۱)، «نجفی‌خواه» و همکاران (۱۳۹۵)،
۲. اقتصادی	۱-۲. هزینه نرم‌افزار و سخت‌افزار ۲-۲. هزینه مشاوره طرح/ پروژه ۳-۲. هزینه‌های پنهان مانند زمان مجریان، مطالعه مقدماتی، و آموزش ۴-۲. هزینه واکاوی وضع موجود ۵-۲. منابع مالی لازم برای پیاده‌سازی سیستم ۶-۲. هزینه نیروی انسانی/ تیم اجرایی ۷-۲. هزینه آموزش (چنانچه به کار بستن سیستم، مستلزم آموزش باشد) ۸-۲. هزینه تجهیزات و اسباب و اثاثیه تازه برای مکان انجام پروژه (رایانه، میز، صندلی، و....) ۹-۲. هزینه نرم‌افزار و پشتیبانی، و هزینه منابع و پایگاه‌های اطلاعاتی الکترونیکی	(همتی و همکاران: ۱۳۹۳)، «موخری» و «رئی» (۲۰۱۷)، «بوس» و همکاران (۲۰۱۴)، «یانگ» و همکاران (۲۰۱۲)، «جواهری» (۱۳۸۳)، «پورفرج» و همکاران (۱۳۹۱)، «نجفی‌خواه» و همکاران (۱۳۹۵)،
۳. زمان‌بندی	۱-۳. زمان توسعه یک سیستم ۲-۳. امکان شناور ساختن پروژه ۳-۳. روشن کردن الزامی یا دلخواه بودن زمان تحویل پروژه	پاور و همکاران (۱۹۹۰)، «موخری» و «رئی» (۲۰۱۷)، «بوس» و همکاران (۲۰۱۴)
۴. عملیاتی	۱-۴. پذیرا بودن بستر سازمان، تشکیلات، و محیط برای پیاده‌سازی پروژه ۲-۴. همخوانی سازمانی و فرهنگی ۳-۴. تأثیر سیستم پیشنهادی روی کارکنان و مدیران ۴-۴. مناسب بودن طراحی محیط کار برای انجام پروژه ۵-۴. وظایف مدیران و کاربران نهایی ۶-۴. مقاومت کاربران نهایی ۷-۴. احساس کاربران نهایی دربارهٔ انجام یک فرآیند جدید ۸-۴. تحت تأثیر قرار گرفتن محیط کاری و اندازه این تأثیر	(شرف: ۱۳۸۶، نقل از سالاری: ۱۳۹۵)، «موخری» و «رئی» (۲۰۱۷)، «بوس» و همکاران (۲۰۱۴)، «ادیب‌منش» و همکاران (۱۳۹۰)، شید (۲۰۱۳)

ابعاد	معیارها/ مقیاس‌ها	منابع
	۴-۹. سود سیستم پیشنهادی برای سازمان	
۵. عوامل انسانی	۵-۱. واکنش مردم به کاربرد یک سیستم جدید ۵-۲. دیرکرد یا اشکال در ساخت سیستم پیشنهادی بر اثر واکنش مردم/ کاربران	پاور و همکاران (۱۹۹۰)
۶. قانونی حقوقی	۶-۱. امکان تناقض سیستم پیشنهادی با الزام‌های قانونی ۶-۲. قوانین سرمایه‌گذاری خارجی ۶-۳. قوانین اعطای حق مالکیت یا کنترل دارایی عمومی یا قوانین مسئولیت تحویل «موخری» و «ری» (۲۰۱۷)، سرویس‌های عمومی به شخص ثالث ۶-۴. چگونگی برخورد قانونی با منابع درآمد	(ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور)، «بوس» و همکاران (۲۰۱۴)، (ppp-certification.com)
۷. فرهنگی	۷-۱. پذیرش فرهنگی سیستم پیشنهادی در جو سازمانی	(ویتنی: ۲۰۰۷، نقل از صراف‌زاده و علیپور: ۱۳۹۰)
۸. منابع	۷-۱. منابع انجام پروژه ۷-۲. امکانات انجام پروژه ۷-۳. امکان استفاده از یارانه دولتی	(ppp-virtualservices.com) «یانگ» و همکاران (۲۰۱۲)، «جواهری» (۱۳۸۳)، «نجفی-خواه» و همکاران (۱۳۹۵)
۹. بازاریابی	۹-۱. مشتریان محصول پس از ساخت آن ۹-۲. آزمایش محصول پیش از عرضه به بازار	«جواهری» (۱۳۸۳)، «پورفرج» و همکاران (۱۳۹۱)، (ppp-evirtualservices.com)

بر اساس اطلاعات جدول ۴-۱، پرسش‌نامه‌ای تهیه شد که در آن ابعاد، معیارها و مقیاس‌ها با توجه به ساخت پیکره زبانی تعیین شد؛ به عنوان مثال، در بعد فنی معیار «مهارت‌ها و تخصص‌های فنی و دانش لازم جهت توسعه/به کار بستن/ ساخت سیستم»، با توجه به ساخت پیکره مشخص شد، به این صورت که مشخص شد منظور از تخصص‌های فنی لازم در این زمینه چیست. لازم به ذکر است که از آنجایی که هدف، امکان‌سنجی ساخت پیکره با استفاده از داده‌های موجود در ایرانداک است و داده‌های موجود، نوعاً داده‌های نوشتاری هستند و نه گفتاری، بنابراین مبنای طراحی پرسش‌نامه، مباحث فنی، اقتصادی و مربوط به ساخت پیکره متنی است تا در نهایت بتوان مدل امکان‌سنجی ساخت پیکره متنی را ارائه داد.

این پرسش‌نامه سه بخش دارد. در بخش نخست، دیدگاه پاسخ‌گویان درباره مؤلفه‌های ساخت پیکره متنی خواسته می‌شود. این مؤلفه‌ها از بررسی پژوهش‌های پیشین به‌دست آمده‌اند. در بخش دوم از پاسخ‌گویان درخواست می‌شود که مؤلفه‌های کلیدی امکان‌سنجی ساخت پیکره متنی را که در بخش نخست نبوده‌اند، پیشنهاد کنند. مؤلفه‌های پیشنهادی، جمع‌بندی می‌شوند و در پرسش‌نامه‌ای دیگر برای دریافت دیدگاه پاسخ‌گویان ارائه می‌شود. در بخش سوم نیز دیدگاه پاسخ‌گویان درباره ابزارهای لازم برای ساخت پیکره متنی پرسیده می‌شود. پرسش‌نامه مذکور در جدول ۴-۲، قابل مشاهده است:

جدول ۴-۲: پرسش‌نامه اولیه ساخت پیکره متنی

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

با بودن هر یک از مؤلفه‌های زیر					در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
کاملاً موافقم ۵	تا اندازه-ای موافقم ۴	مطمئن نیستم ۳	تا اندازه-ای مخالفم ۲	کاملاً مخالفم ۱	
۱. بُعد فنی					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۱-۱. بررسی نیاز به زبان‌شناس همگانی/عمومی در تیم ساخت پیکره ..
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره
۴-۱. بررسی نیاز به متخصص دستور و زبان فارسی در تیم ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۵-۱. بررسی کاربرد نرم‌افزارهای آماده برای ساخت شکل کلی پیکره ..
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۶-۱. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره
۷-۱. طراحی زیربنای پیکره (ساختار و شکل کلی)، به عنوان گامی از ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۸-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۹-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۱۰-۱. حاشیه‌نویسی، به عنوان گامی از ساخت پیکره
۲. بُعد اقتصادی					

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

کاملاً موافقم ۵	تا اندازه-ای موافقم ۴	مطمئن نیستم ۳	تا اندازه-ای مخالفم ۲	کاملاً مخالفم ۱	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
☐	☐	☐	☐	☐	۱-۲. برآورد هزینه نیروی انسانی
☐	☐	☐	☐	☐	۲-۲. برآورد هزینه مطالعات مقدماتی برای ساخت پیکره
☐	☐	☐	☐	☐	۳-۲. برآورد نیاز به هزینه مشاوره با متخصصان ساخت پیکره
☐	☐	☐	☐	☐	۴-۲. برآورد هزینه آموزش کاربران
☐	☐	☐	☐	☐	۵-۲. برآورد هزینه حاشیه‌نویسی دستی
☐	☐	☐	☐	☐	۶-۲. برآورد هزینه خرید نرم‌افزارهای لازم
☐	☐	☐	☐	☐	۷-۲. برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و.....)
☐	☐	☐	☐	☐	۳. بُعد زمان‌بندی
☐	☐	☐	☐	☐	۱-۳. برآورد زمان‌بندی ساخت پیکره
☐	☐	☐	☐	☐	۴. بُعد قانونی/ حقوقی
☐	☐	☐	☐	☐	۱-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره
☐	☐	☐	☐	☐	۲-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره
☐	☐	☐	☐	☐	۳-۴. بررسی حقوق مادی پیکره (پس از ساخت)
☐	☐	☐	☐	☐	۵. بُعد عملیاتی
☐	☐	☐	☐	☐	۱-۵. بررسی پذیرش سازمانی در کاربرد پیکره
☐	☐	☐	☐	☐	۲-۵. بررسی تأثیر پیکره بر سازوکارهای درون‌سازمانی
☐	☐	☐	☐	☐	۳-۵. بررسی دستاوردهای سازمان با ساخت پیکره
☐	☐	☐	☐	☐	۴-۵. پیش‌بینی وظایف کاربران در ساخت پیکره

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

با بودن هر یک از مولفه‌های زیر					در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
کاملاً موافقم ۵	تا اندازه-۱ موافقم ۴	مطمئن نیستم ۳	تا اندازه-۱ مخالفم ۲	کاملاً مخالفم ۱	
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۵-۵. پیش‌بینی وظایف مدیران سازمان در ساخت پیکره.....
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۶. بُعد تأمین منابع
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۶-۱. پیش‌بینی منابع تأمین مالی برای ساخت پیکره.....
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۷. بُعد بازاریابی
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۷-۱. شناخت کاربران پیکره.....
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	۷-۲. آزمون عملکرد پیکره پیش از عرضه به بازار.....

بخش دوم: پیشنهادهای شما درباره مؤلفه‌های کلیدی امکان‌سنجی ساخت پیکره متنی که در بخش نخست نبودند.

۱. بُعد فنی	
۲. بُعد اقتصادی	
۳. بُعد قانونی / حقوقی	
۴. بُعد زمان‌بندی	
۵. بُعد عملیاتی	
۶. بُعد تأمین منابع	

۷. بُعد بازاریابی	
-------------------	--

بخش سوم: دیدگاه شما درباره ابزارهای لازم برای ساخت پیکره متنی.

۱. کاربرد کدام یک از ابزارهای پردازش زبان طبیعی را در ساخت پیکره متنی ضروری می‌دانید؟
(می‌توانید چند گزینه را نیز همزمان انتخاب کنید.)

- ☐ ابزار جداساز (Tokenizer)
- ☐ ابزار نرمال‌سازی متن / پیش‌پردازش متن (Normalizer)
- ☐ پارسر (Parser)
- ☐ ابزار ریشه‌یابی (Stemming)
- ☐ ابزار برچسب‌گذاری اجزای واژگانی کلام (POS tagging)
- ☐ ابزار تشخیص‌دهنده گروه‌های نحوی (Chunker)
- ☐ خطایاب (Error detector)
- ☐ ابزارهای دیگر (خواهشمند است نام فارسی و انگلیسی آن‌ها را بنویسید):

۲. کاربرد کدام یک از حاشیه‌نویسی‌ها را در ساخت پیکره متنی ضروری می‌دانید؟
(می‌توانید چند گزینه را نیز همزمان انتخاب کنید.)

- ☐ حاشیه‌نویسی نحوی (Syntactic annotation)
- ☐ حاشیه‌نویسی کاربردی (Pragmatic annotation)
- ☐ حاشیه‌نویسی گفتمانی (Discourse annotation)
- ☐ حاشیه‌نویسی سبکی (Stylistic annotation)
- ☐ حاشیه‌نویسی واژگانی (Lexical annotation)
- ☐ حاشیه‌نویسی‌های دیگر (خواهشمند است نام فارسی و انگلیسی آن‌ها را بنویسید):

پیش از ارائه پرسش‌نامه به پاسخ‌گویان، برای اطمینان از روایی پرسش‌نامه، از سه متخصص حوزه زبان-شناسی رایانشی خواسته شد تا پرسش‌نامه را مطالعه کنند و نظراتشان را در قالب پاسخ به سؤالاتی که در جدول ۴-۳، قابل مشاهده است، وارد نمایند.

جدول ۴-۳: نحوه اظهار نظر متخصصان درباره کلیت پرسش‌نامه امکان‌سنجی ساخت پیکره متنی

۱. آیا گویه‌ای وجود دارد که پاسخ‌گویان را ناراحت کند؟	
شماره پرسش	توضیح یا پیشنهاد پرسش جایگزین
۲. آیا گویه‌ای وجود دارد که درک آن، مستلزم دو یا چند بار خواندن باشد؟	
شماره پرسش	توضیح یا پیشنهاد پرسش جایگزین
۳. آیا گویه‌ای وجود دارد که باعث سوء برداشت یا تفسیر نادرست از طرف پاسخ‌گویان شود؟	
شماره پرسش	توضیح یا پیشنهاد پرسش جایگزین
۴. آیا گویه‌ای وجود دارد که خواندن آن دشوار به نظر برسد؟	
شماره پرسش	توضیح یا پیشنهاد پرسش جایگزین
۵. آیا گویه‌های مربوط به هر یک از ابعاد، خسته‌کننده به نظر نمی‌رسند؟	
شماره پرسش	توضیح یا پیشنهاد
۶. آیا پاسخ‌هایی که برای هر گویه در نظر گرفته شده‌اند، کافی به نظر می‌رسند؟	
شماره پرسش	توضیح یا پیشنهاد

پاسخ متخصصان به این بخش از پژوهش (اظهارنظر درباره کلیت پرسش‌نامه)، در قالب جدول ۴-۴، خلاصه شده است.

جدول ۴-۴: پاسخ متخصصان زبان‌شناسی رایانشی درباره روایی پرسش‌نامه اولیه ساخت پیکره متنی

شماره پرسش	دیدگاه درباره پرسش‌ها	متخصص نخست	متخصص دوم	متخصص سوم
۱-۱				
۲-۱				
۳-۱	ناراحت کردن پاسخ-گویان		باید به جای «متخصص رایانه‌ای» عبارت «متخصص فناوری اطلاعات» استفاده شود	
۴-۱	تفسیر نادرست پرسش-ها			این پرسش با ۱-۱ همپوشانی دارد زیرا زبان‌شناسان عمومی قاعدتاً می‌بایست به دستور زبان فارسی مسلط باشند و توانایی تحلیل جمله‌های فارسی و ساخت درونی واژه‌ها را داشته باشند. لذا، ممکن است پرسش ۴-۱ به این تصور بینجامد که منظور شما فارغ‌التحصیلان زبان و ادبیات فارسی است که البته به تجربه و با آگاهی از مواد درسی ایشان می‌دانیم حضورشان در پروژه‌های پیکره اگر مضر نباشد، مفید هم نیست.
۵-۱	تفسیر نادرست پرسش-ها	بررسی مزایا و معایب استفاده از نرم‌افزارهای آماده در مقابل نرم‌افزار جدید (مثلاً هزینه).		
۶-۱				
۷-۱	نیاز به چندبار خواندن پرسش‌ها		این عبارت وارد شود: طراحی پیکره (چارچوب و مولفه‌های آن)، به عنوان گامی از ساخت پیکره	
۸-۱	تفسیر نادرست پرسش-ها	علاوه بر نوع متن، به حجم پیکره نیز اشاره شود.		

۹-۱				
۱۰-۱	نیاز به چندبار خواندن پرسش‌ها		این عبارت وارد شود: حاشیه‌نویسی (برچسب گذاری)، به عنوان گامی از ساخت پیکره	
۱-۲	تفسیر نادرست پرسش‌ها		این پرسش بیش از حد کلی به نظر می‌آید و می‌تواند به چند پرسش کوچک‌تر تقسیم شود که هر کدام از آنها جنبه‌ای از هزینه‌های مربوط به نیروی انسانی را هدف می‌گیرند.	
۲-۲	دشواری خواندن پرسش‌ها	این مورد به نظر زیر مجموعه‌ای است از مورد ۲-۱. شاید بهتر باشد که مورد ۲-۱ به سایر هزینه‌های نیروی انسانی تغییر یابد.		
۳-۲	دشواری خواندن پرسش‌ها	این مورد به نظر زیر مجموعه‌ای است از مورد ۲-۱. شاید بهتر باشد که مورد ۲-۱ به سایر هزینه‌های نیروی انسانی تغییر یابد		
۴-۲	دشواری خواندن پرسش‌ها	این مورد به نظر زیر مجموعه‌ای است از مورد ۲-۱. شاید بهتر باشد که مورد ۲-۱ به سایر هزینه‌های نیروی انسانی تغییر یابد		
۵-۲	دشواری خواندن پرسش‌ها	این مورد به نظر زیر مجموعه‌ای است از مورد ۲-۱. شاید بهتر باشد که مورد ۲-۱ به سایر هزینه‌های نیروی انسانی تغییر یابد		
۶-۲				

۷-۲				
۱-۳	دشواری خواندن پرسش‌ها	زمان‌بندی با توجه به امکانات میسر است به طریقی باید در این سوال در نظر گرفته شود.		
۱-۴				
۲-۴				
۳-۴				
۱-۵	دشواری خواندن پرسش‌ها		این پرسش به توضیح بیشتری نیاز دارد.	
۲-۵				
۳-۵				
۴-۵				
۵-۵				
۱-۶	تفسیر نادرست پرسش‌ها	جواب این سوال واضح است، نیازی به طرح این سوال نیست.		
۱-۷				
۲-۷				
بخش دوم				
بخش سوم	تفسیر نادرست پرسش‌ها	پاسخ به این سوالات نیازمند تعیین نوع و هدف پیکره‌ای است که قصد ایجاد آن را داریم. با توجه به نامعلوم بودن نوع پیکره، سوالات مطرح شده در این بخش ناواضح به نظر می‌رسند.	به نظر می‌رسد این سوال باعث شود که متخصصان ابزارهای پایه لیست شده را علامت بزنند. در صورتیکه به نظر اینجانب بعضی از پیکره‌ها نیازی به هیچ ابزاری ندارند.	

در این قسمت به بررسی نظرات متخصصان و نحوه اعمال نظرات ایشان و دلایل عدم پذیرش برخی از نظرات پرداخته می‌شود. خلاصه اعمال نظرات متخصصان در جدول ۴-۵، قابل مشاهده است.

جدول ۴-۵: روش اعمال یا رد نظر متخصصان زبان‌شناسی رایانشی درباره روایی پرسش‌نامه اولیه ساخت پیکره متنی

شماره پرسش	متخصص نخست	متخصص دوم	متخصص سوم	اعمال نظر متخصص / ارائه دلیل رد نظر متخصص
۳-۱		باید به جای «متخصص رایانه‌ای» عبارت «متخصص فناوری اطلاعات» استفاده شود		لزوماً متخصص فناوری اطلاعات مسلط به کار برنامه‌نویسی رایانه‌ای نیست. حال آنکه در کار ساخت پیکره حتماً نیازمند کسی هستیم تا به برنامه‌نویسی تسلط داشته باشد تا بتواند در کار ساخت شکل کلی پیکره و چگونگی وارد کردن اطلاعات و.....همکاری کند.
۴-۱			این پرسش با ۱-۱ همپوشانی دارد زیرا زبان‌شناسان عمومی قاعدتاً می‌بایست به دستور زبان فارسی مسلط باشند و توانایی تحلیل جمله‌های فارسی و ساخت درونی واژه‌ها را داشته باشند. لذا، ممکن است پرسش ۴-۱ به این تصور بینجامد که منظور شما فارغ‌التحصیلان زبان و ادبیات فارسی است که البته به تجربه و با آگاهی از مواد درسی ایشان می‌دانیم حضورشان در پروژه‌های پیکره اگر مضر نباشد، مفید هم نیست.	نظر متخصص کاملاً منطقی است، بنابراین، این پرسش در نسخه ویرایش شده، حذف شد
۵-۱	بررسی مزایا و معایب استفاده از نرم‌افزارهای آماده در مقابل نرم‌افزار جدید (مثلاً هزینه).			این پرسش در نسخه ویرایش شده به صورت زیر اصلاح شد: بررسی کاربرد نرم‌افزارهای

				آماده در مقایسه با تولید نرم‌افزار جدید برای ساخت شکل کلی پیکره
۷-۱	این عبارت وارد شود: طراحی پیکره (چارچوب و مولفه-های آن)، به عنوان گامی از ساخت پیکره		این پرسش در نسخه ویرایش شده به صورت زیر اصلاح شد: طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره	
۸-۱	علاوه بر نوع متن، به حجم پیکره نیز اشاره شود.		این پرسش، در نسخه ویرایش شده وارد شد.	
۱۰-۱	این عبارت وارد شود: حاشیه-نویسی (برچسب گذاری)، به عنوان گامی از ساخت پیکره		این پرسش در نسخه ویرایش شده به صورت زیر اصلاح شد: حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره	
۱-۲			این پرسش بیش از حد کلی به نظر می‌آید و می‌تواند به چند پرسش کوچک‌تر تقسیم شود که هر کدام از آنها جنبه‌ای از هزینه‌های مربوط به نیروی انسانی را هدف می‌گیرند. این پرسش به نظر متخصص دوم در مورد این پرسش به نظر متخصص اول در موارد ۲-۲ تا ۵-۲، نزدیک است. در نسخه ویرایش شده نظرات دو متخصص به این صورت اعمال شد که مورد ۱-۲ حذف شد، زیرا در واقع در موارد بعدی مشخص شده است که منظور از نیروی انسانی همان متخصصان ساخت پیکره، برنامه‌نویس رایانه‌ای، آموزش کاربران، حاشیه نویسان دستی و... است. مورد ۲-۲ هم به این دلیل حذف شد که در واقع مطالعه مقدماتی	

را متخصصان ساخت پیکره انجام خواهند داد که در هزینه‌ها در نظر گرفته شده است.				
			این مورد به نظر زیر مجموعه‌ای است از مورد ۲-۱. شاید بهتر باشد که مورد ۲-۱ به سایر هزینه‌های نیروی انسانی تغییر یابد.	۲-۲ تا ۵-۲
در بعد اقتصادی و بخش‌های دیگر، امکانات دیده شده است			زمان‌بندی با توجه به امکانات میسر است به طریقی باید در این سوال در نظر گرفته شود.	۳-۱
این پرسش در نسخه ویرایش شده به صورت زیر اصلاح شده است: بررسی پذیرش و آمادگی سازمان در کاربرد پیکره	این پرسش به توضیح بیشتری نیاز دارد.			۵-۱
خیر، پاسخ این سوال واضح نیست. زیرا باید مشخص شود تامین مالی از چه منابعی و چگونه انجام خواهد شد. به عنوان مثال، چه مقدار از هزینه را سازمان تقبل می‌کند، چه مقدار از سازمان/سازمان‌های دیگر یا حتی کاربران نهایی.			جواب این سوال واضح است، نیازی به طرح این سوال نیست.	۶-۱
از آنجایی که ابزارهای پردازش زبان ذکر شده در این بخش و انواع حاشیه‌نویسی‌ها شناخته شده هستند، بنابراین این سوال به این صورت اصلاح شد، که غیر از این موارد اگر موارد دیگری وجود دارد، پاسخ‌گو وارد کند. ضمناً صورت سوال هم به صورت زیر اصلاح شد:		به نظر می‌رسد این سوال باعث شود که متخصصان ابزارهای پایه لیست شده را علامت بزنند. در صورتیکه به نظر اینجانب بعضی از پیکره‌ها نیازی به هیچ ابزاری ندارند.	پاسخ به این سوالات نیازمند تعیین نوع و هدف پیکره‌ای است که قصد ایجاد آن را داریم. با توجه به نامعلوم بودن نوع پیکره، سوالات مطرح شده در این بخش ناواضح به نظر می‌رسند.	بخش سوم

- به جز ابزارهای پردازش زبان طبیعی زیر، روی هم رفته چه ابزارهای دیگری را در ساخت پیکره متنی دارای کاربرد می‌دانید؟ - به جز حاشیه‌نویسی‌های زیر، روی هم رفته چه حاشیه‌نویسی‌هایی را در ساخت پیکره متنی دارای کاربرد می‌دانید؟				
--	--	--	--	--

در نهایت پرسش‌نامه به صورت جدول ۴-۶، اصلاح شد.

جدول ۴-۶: پرسش‌نامه ساخت پیکره متنی، پس از ویرایش توسط متخصصان

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

کاملاً مخالفم ۱	تا اندازه ای مخالفم ۲	مطمئن نیستم ۳	تا اندازه ای موافقم ۴	کاملاً موافقم ۵	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
--------------------	--------------------------	------------------	--------------------------	--------------------	---

۱. بُعد فنی

- ۱-۱. بررسی نیاز به زبان‌شناس همگانی / عمومی در تیم ساخت پیکره
☐ ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐ ۵
- ۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره
☐ ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐ ۵
- ۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره
☐ ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐ ۵
- ۴-۱. بررسی کاربرد نرم‌افزارهای آماده در مقایسه با تولید نرم‌افزار جدید برای ساخت شکل کلی پیکره
☐ ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐ ۵
- ۵-۱. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره
☐ ۱ ☐ ۲ ☐ ۳ ☐ ۴ ☐ ۵

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

کاملاً مخالفم ۱	تا اندازه- ای مخالفم ۲	مطمئن نیستم ۳	تا اندازه- ای موافقم ۴	کاملاً موافقم ۵	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
					۶-۱. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۷-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۸-۱. تعیین حجم متون، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۹-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره.....
					۱۰-۱. حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	
					۲- بُعد اقتصادی
☐	☐	☐	☐	☐	۱-۲. برآورد هزینه متخصصان ساخت پیکره.....
☐	☐	☐	☐	☐	۲-۲. برآورد هزینه برنامه‌نویسی رایانه.....
☐	☐	☐	☐	☐	۳-۲. برآورد هزینه حاشیه‌نویسی دستی.....
☐	☐	☐	☐	☐	۴-۲. برآورد هزینه آموزش کاربران.....
☐	☐	☐	☐	☐	۵-۲. برآورد هزینه خرید نرم‌افزار.....
☐	☐	☐	☐	☐	۶-۲. برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و.....)
☐	☐	☐	☐	☐	
					۳- بُعد زمان‌بندی
☐	☐	☐	☐	☐	۱-۳. برآورد زمان‌بندی ساخت پیکره.....
					۴- بُعد قانونی / حقوقی
☐	☐	☐	☐	☐	۱-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره.....
☐	☐	☐	☐	☐	۲-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره.....

بخش نخست: دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین به دست آمده‌اند.

کاملاً موافقم ۵	تا اندازه-۱ موافقم ۴	مطمئن نیستم ۳	تا اندازه-۱ مخالفم ۲	کاملاً مخالفم ۱	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
☐	☐	☐	☐	☐	۳-۴. بررسی حقوق مادی پیکره (پس از ساخت).....
☐	☐	☐	☐	☐	۵- بُعد عملیاتی
☐	☐	☐	☐	☐	۱-۵. بررسی پذیرش و آمادگی سازمان در کاربرد پیکره.....
☐	☐	☐	☐	☐	۲-۵. بررسی تأثیر پیکره بر سازوکارهای درون‌سازمانی.....
☐	☐	☐	☐	☐	۳-۵. بررسی دستاوردهای سازمان با ساخت پیکره.....
☐	☐	☐	☐	☐	۴-۵. پیش‌بینی وظایف کاربران در ساخت پیکره.....
☐	☐	☐	☐	☐	۵-۵. پیش‌بینی وظایف مدیران سازمان در ساخت پیکره.....
☐	☐	☐	☐	☐	۶- بُعد تأمین منابع
☐	☐	☐	☐	☐	۱-۶. پیش‌بینی منابع تأمین مالی برای ساخت پیکره.....
☐	☐	☐	☐	☐	۷- بُعد بازاریابی
☐	☐	☐	☐	☐	۱-۷. شناخت کاربران پیکره.....
☐	☐	☐	☐	☐	۲-۷. آزمون عملکرد پیکره پیش از عرضه به بازار.....

بخش دوم: پیشنهادهای شما درباره مؤلفه‌های کلیدی امکان‌سنجی ساخت پیکره متنی که در بخش نخست نبودند.

۱. بُعد فنی	
۲. بُعد اقتصادی	

	۳. بُعد قانونی / حقوقی
	۴. بُعد زمان‌بندی
	۵. بُعد عملیاتی
	۶. بُعد تأمین منابع
	۷. بُعد بازاریابی

بخش سوم: دیدگاه شما درباره ابزارهای لازم برای ساخت پیکره متن

۱. به جز ابزارهای پردازش زبان طبیعی زیر، روی هم‌رفته چه ابزارهای دیگری را در ساخت پیکره متن دارای کاربرد می‌دانید؟
(می‌توانید چند گزینه را نیز همزمان انتخاب کنید).

- ابزار جداساز (Tokenizer)
 - ابزار نرمال‌سازی متن / پیش‌پردازش متن (Normalizer)
 - پارسر (Parser)
 - ابزار ریشه‌یابی (Stemming)
 - ابزار برچسب‌گذاری اجزای واژگانی کلام (POS tagging)
 - ابزار تشخیص‌دهنده گروه‌های نحوی (Chunker)
 - خطایاب (Error detector)
- خواهشمند است نام فارسی و انگلیسی ابزارهای دیگر را بنویسید:

۲. به جز حاشیه‌نویسی‌های زیر، روی هم‌رفته چه حاشیه‌نویسی‌هایی را در ساخت پیکره متن دارای کاربرد می‌دانید؟
(می‌توانید چند گزینه را نیز همزمان انتخاب کنید).

- ♦ حاشیه‌نویسی نحوی (Syntactic annotation)
 - ♦ حاشیه‌نویسی کاربردی (Pragmatic annotation)
 - ♦ حاشیه‌نویسی گفتمانی (Discourse annotation)
 - ♦ حاشیه‌نویسی سبکی (Stylistic annotation)
 - ♦ حاشیه‌نویسی واژگانی (Lexical annotation)
- خواهشمند است نام فارسی و انگلیسی حاشیه‌نویسی‌های دیگر را بنویسید:

۴-۵-۲- دور اول دلفی

پرسش‌نامه در نهایت به ده متخصص موضوعی تحویل داده شد. ۹ نفر از متخصصان، اعضای هیات علمی دانشگاه‌ها یا پژوهشگاه‌های وابسته به وزارت علوم هستند. رشته تحصیلی و زمینه فعالیت متخصصان شامل: زبان‌شناسی همگانی، زبان‌شناسی رایانشی، ریاضی (علوم کامپیوتر)، مهندسی کامپیوتر و اصطلاح-نامه‌شناسی است. اکثر متخصصان، سابقه بیش از ۳ سال فعالیت در حوزه زبان‌شناسی رایانشی / پیکره‌ای دارند.

جدول ۴-۷، درصد فراوانی انتخاب گزینه‌های کاملاً موافقم تا کاملاً مخالفم، مربوط به بخش نخست را نشان می‌دهد.

جدول ۴-۷: درصد فراوانی انتخاب گزینه‌های مربوط به هر گویه در بخش نخست

شماره گویه	درصد فراوانی انتخاب گزینه ۵ (کاملاً موافقم)	درصد فراوانی انتخاب گزینه ۴ (تا اندازه‌ای موافقم)	درصد فراوانی انتخاب گزینه ۳ (مطمئن نیستم)	درصد فراوانی انتخاب گزینه ۲ (تا اندازه‌ای مخالفم)	درصد فراوانی انتخاب گزینه ۱ (کاملاً مخالفم)
۱-۱	٪۷۰	٪۳۰	٪۰	٪۰	٪۰
۲-۱	٪۸۰	٪۱۰	٪۰	٪۰	٪۱۰
۳-۱	٪۸۰	٪۱۰	٪۰	٪۰	٪۱۰
۴-۱	٪۲۰	٪۶۰	٪۱۰	٪۰	٪۱۰
۵-۱	٪۴۰	٪۲۰	٪۴۰	٪۰	٪۰
۶-۱	٪۸۰	٪۲۰	٪۰	٪۰	٪۰

۷-۱	%۹۰	%۱۰	%۰	%۰	%۰
۸-۱	%۹۰	%۱۰	%۰	%۰	%۰
۹-۱	%۶۰	%۴۰	%۰	%۰	%۰
۱۰-۱	%۴۰	%۶۰	%۰	%۰	%۰
۱-۲	%۸۰	%۲۰	%۰	%۰	%۰
۲-۲	%۹۰	%۱۰	%۰	%۰	%۰
۳-۲	%۶۰	%۳۰	%۱۰	%۰	%۰
۴-۲	%۵۰	%۲۰	%۲۰	%۱۰	%۰
۵-۲	%۷۰	%۲۰	%۱۰	%۰	%۰
۶-۲	%۴۰	%۶۰	%۰	%۰	%۰
۱-۳	%۱۰۰	%۰	%۰	%۰	%۰
۱-۴	%۸۰	%۲۰	%۰	%۰	%۰
۲-۴	%۷۰	%۳۰	%۰	%۰	%۰
۳-۴	%۹۰	%۱۰	%۰	%۰	%۰
۱-۵	%۵۰	%۳۰	%۲۰	%۰	%۰
۲-۵	%۳۰	%۲۰	%۴۰	%۱۰	%۰
۳-۵	%۴۰	%۴۰	%۲۰	%۰	%۰
۴-۵	%۲۰	%۳۰	%۴۰	%۱۰	%۰
۵-۵	%۲۰	%۴۰	%۴۰	%۰	%۰
۱-۶	%۱۰۰	%۰	%۰	%۰	%۰
۱-۷	%۷۰	%۲۰	%۱۰	%۰	%۰
۲-۷	%۶۰	%۳۰	%۱۰	%۰	%۰

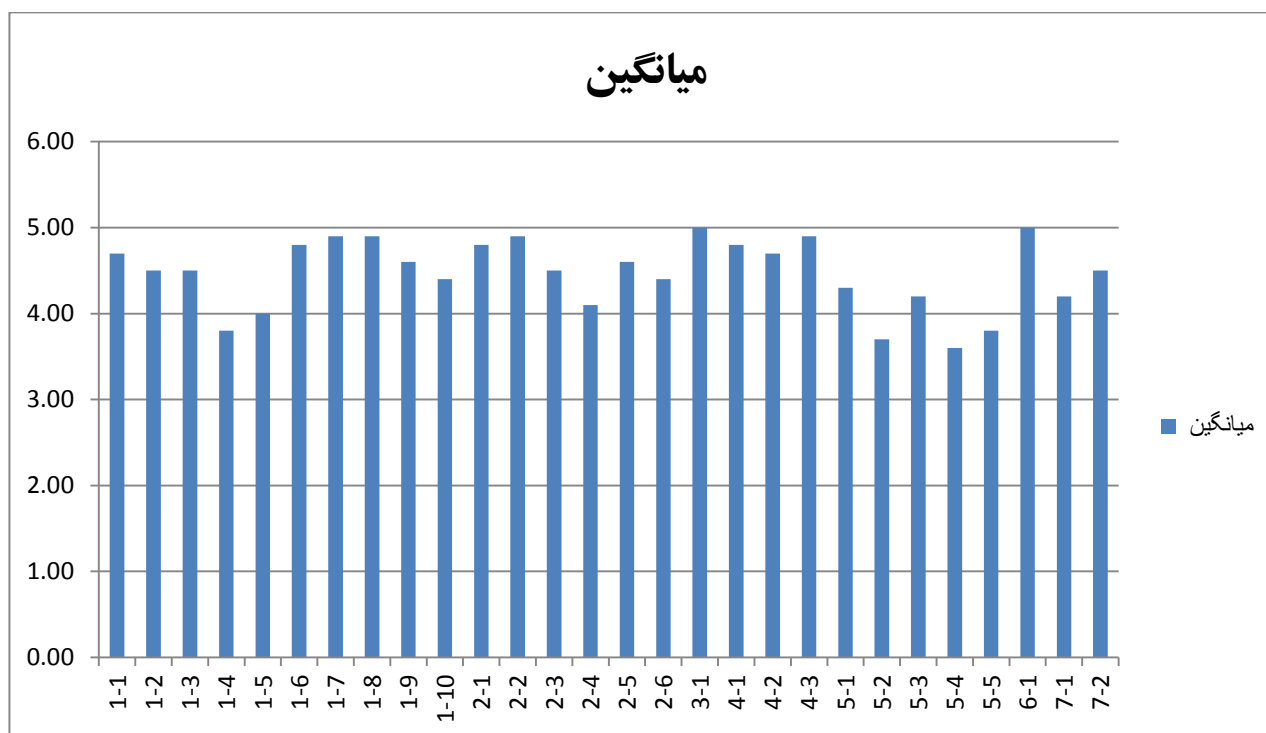
برای تصمیم‌گیری درباره توقف یا ادامه دورهای دلفی می‌توان از معیار میانگین استفاده کرد؛ میانگین، برآیند داده‌ها را نشان می‌دهد و به عنوان معیاری برای تمایل به مرکز، عقیده گروهی اعضای پانل دلفی را نشان می‌دهد.

جدول ۴-۸، و شکل ۴-۳، نمودار میانگین پاسخ به هر گویه، در بخش نخست را نشان می‌دهد.

جدول ۴-۸: میانگین پاسخ به هر گویه در دور اول دلفی

شماره گویه	میانگین پاسخ‌ها
۱-۱	۴,۷۰
۲-۱	۴,۵۰
۳-۱	۴,۵۰
۴-۱	۳,۸۰
۵-۱	۴
۶-۱	۴,۸۰
۷-۱	۴,۹۰
۸-۱	۴,۹۰
۹-۱	۴,۶۰
۱۰-۱	۴,۴۰
۱-۲	۴,۸۰
۲-۲	۴,۹۰
۳-۲	۴,۵۰
۴-۲	۴,۱۰
۵-۲	۴,۶۰
۶-۲	۴,۴۰
۱-۳	۵
۱-۴	۴,۸۰
۲-۴	۴,۷۰
۳-۴	۴,۹۰
۱-۵	۴,۳۰
۲-۵	۳,۷۰
۳-۵	۴,۲۰
۴-۵	۳,۶۰
۵-۵	۳,۸۰
۱-۶	۵

۴,۲۰	۱-۷
۴,۵۰	۲-۷



نمودار ۴-۳: میانگین پاسخ به هر گویه در دور اول دلفی

با توجه به جدول ۴-۸، می‌توان این‌طور نتیجه‌گیری نمود که چنانچه میانگین پاسخ به گویه‌ای ۴ $\leq$ باشد، آن گویه برای امکان‌سنجی لحاظ می‌شود و گویه‌هایی که فاقد این شرایط باشند، از امکان‌سنجی حذف می‌شوند. همانگونه که در جدول ۴-۸ مشخص شده است، میانگین پاسخ به همه گویه‌ها، غیر از گویه‌های ۴-۱، ۲-۵، ۴-۵ و ۵-۵، بالاتر از ۴ است. بنابراین اکثر متخصصان با درج گویه‌ها در امکان-سنجی اتفاق‌نظر دارند و همه گویه‌های موجود در بخش نخست (غیر از گویه‌های ۴-۱، ۲-۵، ۴-۵ و ۵-۵) باید در امکان‌سنجی لحاظ شوند. گویه‌های حذف شده شامل موارد زیر هستند:

گویه ۴-۱: بررسی کاربرد نرم‌افزارهای آماده در مقایسه با تولید نرم‌افزار جدید برای ساخت شکل کلی پیکره

گویه ۲-۵: بررسی تاثیر پیکره بر ساز و کارهای درون‌سازمانی

گویه ۴-۵: پیش‌بینی وظایف کاربران در ساخت پیکره

گویه ۵-۵: پیش‌بینی وظایف مدیران سازمان در ساخت پیکره

جدول ۴-۹، خلاصه نظرات پاسخ‌گویان در بخش دوم (درباره مؤلفه‌های کلیدی امکان‌سنجی ساخت پیکره متنی که در بخش نخست نبودند) را نشان می‌دهد.

جدول ۴-۹: نظر پاسخ‌گویان درباره سایر مؤلفه‌های احتمالی دیگر در ساخت پیکره

- استفاده از نرم‌افزار واسطه (interface) برای کار کردن کاربران با پیکره؛ زیرا کار کردن با نرم‌افزارهای مربوط به پیکره در ایران آسان نیست (به اصطلاح user friendly نیستند). علاوه بر ساخت پیکره، اگر امکان سرمایه‌گذاری در این بخش وجود داشته باشد، بسیار عالی است. - در نظر گرفتن نیاز به پیکره‌ای مناسب برای استفاده در فرهنگ‌نگاری دوزبانه. - استفاده از پژوهشگرانی که تجربه ساخت پیکره دارند در تیم طراحی یا به‌رمندی از مشاوره آن‌ها. - بررسی امکان اتصال و سازگاری آن با سامانه‌های موجود در ایرانداک.	۱. بُعد فنی
- در نظر گرفتن هزینه سرور و پشتیبانی از آن. - در نظر گرفتن هزینه ساخت ابزار جمع‌آوری متن، حاشیه‌نویسی پیکره و هزینه ساخت ابزار کار با داده‌های پیکره هنگام برآورد هزینه برنامه‌نویس رایانه‌ای. - تصمیم‌گیری درباره نحوه دسترسی کاربران به آن: رایگان یا با پرداخت هزینه. - برآورد هزینه به روز کردن داده‌های پیکره در پیکره‌های پیش‌گر/ به‌روزشونده.	۲. بُعد اقتصادی
- تنظیم اجازه‌نامه برای اخذ داده‌های گفتاری از اهل زبان	۳. بُعد قانونی/ حقوقی
- این بخش به بخش‌های کوچک‌تر تقسیم شود؛ پیشنهاد می‌شود زمان‌بندی را فازبندی کنید: زمان‌بندی فاز پژوهشی، زمان‌بندی ساخت ابزار حاشیه‌نویسی، زمان‌بندی حاشیه‌نویسی، زمان‌بندی ساخت ابزار کار با پیکره و.....	۴. بُعد زمان‌بندی
	۵. بُعد عملیاتی

۶. بُعد تأمین منابع	- بررسی امکان تعریف طرح پژوهشی در قالب طرح مورد حمایت از سوی صندوق حمایت از پژوهشگران یا ستاد توسعه علوم شناختی. در این صورت بخشی از تأمین مالی توسط آن‌ها انجام خواهد گرفت.
۷. بُعد بازاریابی	- برنامه‌ریزی نحوه عرضه پیکره - بررسی شیوه استفاده از آن در سامانه‌های خدماتی ایرانداک (به منظور درآمدزایی)

نکاتی که پاسخ‌گویان در بخش سوم ذکر کرده‌اند در قالب جدول ۴-۱۰، قابل مشاهده است.

جدول ۴-۱۰: فهرست سایر ابزارهای پردازش زبان طبیعی و انواع حاشیه‌نویسی

۱- ابزارهای پردازش زبان طبیعی	- ابزار حاشیه‌نویسی پیکره - ابزار شناسایی موجودیت‌های نامدار (named entity recognition) - بسته به نوع کاربرد پیکره (خلاصه‌سازی متن، استخراج کلیدواژه، استخراج نوع متن (مهندسی، علوم انسانی و...)) می‌توان از ابزارهای خاص ذکر شده در این بخش استفاده کرد. - خزش گر (crawler)
۲- حاشیه‌نویسی	- از آنجایی که ظاهراً ساخت پیکره علمی مدنظر است، تعیین فراداده هم مهم است (موضوع متن، جنسیت نویسنده، سال انتشار و....) - حاشیه‌نویسی (نشانه‌گذاری) معنایی (semantic annotation) - در مورد حاشیه‌نویسی گفتمانی، مشخصاً حاشیه‌نویسی حرکت‌های بلاغی (rhetorical moves) چکیده‌ها پیشنهاد می‌شود. کاربرد این نوع حاشیه‌نویسی می‌تواند ارزیابی خودکار صحت چکیده‌نویسی باشد. - برچسب‌گذاری نقش‌های معنایی (semantic role labeling) - حاشیه‌نویسی استعاره (metaphor annotation)

--	--

۴-۵-۳- دور دوم دلفی

در این بخش، نخست به بررسی نظرات متخصصان (مطرح شده در بخش دوم پرسش‌نامه) و نحوه اعمال نظرات آن‌ها در پرسش‌نامه دور دوم دلفی و ارائه دلایل عدم پذیرش برخی از نظرات پرداخته می‌شود. سپس، نظرات آن‌ها طی پرسش‌نامه دیگری در اختیار همان متخصصان قرار خواهد گرفت تا درباره گویه‌های مطرح‌شده نظر خود را با انتخاب گزینه‌ها، بیان کنند.

خلاصه نحوه اعمال نظرات متخصصان در دور دوم دلفی در جدول ۴-۱۱، قابل مشاهده است.

جدول ۴-۱۱: روش اعمال یا رد نظر متخصصان درباره دیدگاه‌های مطرح‌شده در پرسش‌نامه دور اول دلفی

ابعاد	نظرات متخصصان	اعمال نظر متخصص / ارائه دلیل رد نظر متخصص
۱. بُعد فنی	۱. استفاده از نرم‌افزار واسطه (interface) برای کار کردن کاربران با پیکره؛ زیرا کارکردن با نرم-افزارهای مربوط به پیکره در ایران آسان نیست (به اصطلاح user friendly نیستند). علاوه بر ساخت پیکره، اگر امکان سرمایه‌گذاری در این بخش وجود داشته باشد، بسیار عالی است.	۱. این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود.
	۲. در نظر گرفتن نیاز به پیکره‌ای مناسب برای استفاده در فرهنگ‌نگاری دوزبانه.	۲. این نکته به کاربرد پیکره مناسب مربوط می‌شود؛ در واقع می‌توان گفت، یک نیاز مطرح شده است و نه مسئله‌ای که باید در بخش فنی مدنظر قرار گیرد.
	۳. استفاده از پژوهشگرانی که تجربه ساخت پیکره دارند در تیم طراحی یا به‌رمندی از مشاوره آن‌ها.	۳. این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود.

۴. این نکته را می‌توان هنگام امکان‌سنجی و عملیاتی کردن ساخت پیکره برای ایرانداک مدنظر قرار داد.	۴. بررسی امکان اتصال و سازگاری آن با سامانه‌های موجود در ایرانداک.	
۱. این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود. ۲. از میان این نکات، تنها «هزینه جمع‌آوری متن» و «هزینه ساخت ابزار کار با داده‌های پیکره» در پرسش‌نامه دور دوم دلفی لحاظ می‌شود. زیرا هزینه حاشیه‌نویسی پیکره در پرسش‌نامه دور اول دلفی درج شده است. ۳. این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود. ۴. این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود.	۱. برآورد هزینه سرور و پشتیبانی از آن. ۲. در نظر گرفتن هزینه ساخت ابزار جمع‌آوری متن، حاشیه‌نویسی پیکره و هزینه ساخت ابزار کار با داده‌های پیکره هنگام برآورد هزینه برنامه‌نویس رایانه‌ای. ۳. تصمیم‌گیری درباره نحوه دسترسی کاربران به آن: رایگان یا با پرداخت هزینه. ۴. برآورد هزینه به روز کردن داده‌های پیکره در پیکره‌های پیشگر/ به روزشونده.	۲. بُعد اقتصادی
همانگونه که در توضیح ابتدای پرسش‌نامه هم ذکر شده است، هدف، امکان‌سنجی ساخت پیکره متنی است؛ بنابراین، نیازی به اخذ اجازه‌نامه برای دریافت داده‌های گفتاری نیست، زیرا اساساً داده گفتاری وجود ندارد.	- تنظیم اجازه‌نامه برای اخذ داده‌های گفتاری از گویشوران	۳. بُعد قانونی/ حقوقی
این نکته در پرسش‌نامه دور دوم دلفی لحاظ می‌شود.	- این بخش به بخش‌های کوچک‌تر تقسیم شود؛ پیشنهاد می‌شود زمان‌بندی را فازبندی کنید: زمان-بندی فاز پژوهشی، زمان‌بندی ساخت ابزار حاشیه‌نویسی، زمان‌بندی حاشیه‌نویسی، زمان‌بندی ساخت	۴. بُعد زمان-بندی

	ابزار کار با پیکره و.....	
۵. بُعد تأمین منابع	- بررسی امکان تعریف طرح پژوهشی در قالب طرح مورد حمایت از سوی صندوق حمایت از پژوهشگران یا ستاد توسعه علوم شناختی. در این صورت بخشی از تأمین مالی توسط آن‌ها انجام خواهد گرفت.	این مسئله هنگام عملیاتی کردن طرح ساخت پیکره مدنظر قرار خواهد گرفت و نیازی به بررسی این مورد در امکان‌سنجی ساخت پیکره متنی نیست؛ زیرا این امکان‌سنجی یک طرح کلی برای ساخت هر نوع پیکره متنی است.
۶. بُعد بازاریابی	۱. برنامه‌ریزی نحوه عرضه پیکره ۲. بررسی شیوه استفاده از آن در سامانه‌های خدماتی ایرانداک (به منظور درآمدزایی)	۱. این مسئله به بخش بعد اقتصادی مربوط می‌شود که آیا به عنوان مثال، رایگان عرضه شود یا با دریافت هزینه باشد. ۲. این نکته را می‌توان هنگام امکان‌سنجی و عملیاتی کردن ساخت پیکره برای ایرانداک مدنظر قرار داد.

بر اساس نتایج جدول ۴-۱۱، پرسش‌نامه دیگری تهیه شد که در جدول ۴-۱۲، قابل مشاهده است.

جدول ۴-۱۲: پرسش‌نامه دور دوم دلفی (به مشورت گذاردن نظرات متخصصان)

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی نظرات متخصصان به دست آمده است.

با بودن هر یک از مولفه‌های زیر				
کاملاً موافقم	تا اندازه-۱ موافقم	مطمئن نیستم	تا اندازه-۱ مخالفم	کاملاً مخالفم
۵	۴	۳	۲	۱
در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟				

۱. بُعد فنی

۱-۱. بررسی نیاز به رابط کاربر (User interface) برای کار

کردن کاربران با پیکره □۵ □۴ □۳ □۲ □۱

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی نظرات متخصصان به دست آمده است.

کاملاً موافقم ۵	تا اندازه-۱ موافقم ۴	مطمئن نیستم ۳	تا اندازه-۱ مخالقم ۲	کاملاً مخالفم ۱	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
☐	☐	☐	☐	☐	۲-۱. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند.....
☐	☐	☐	☐	☐	۲-۲. بُعد اقتصادی
☐	☐	☐	☐	☐	۱-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و...).....
☐	☐	☐	☐	☐	۲-۲. برآورد هزینه نگهداری زیرساخت (سرور، شبکه، و...).....
☐	☐	☐	☐	☐	۳-۲. برآورد هزینه ساخت ابزار گردآوری متن.....
☐	☐	☐	☐	☐	۴-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره.....
☐	☐	☐	☐	☐	۵-۲. تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه).....
☐	☐	☐	☐	☐	۶-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پیشگر / پژوهشونده.....
☐	☐	☐	☐	☐	۳- بُعد زمان‌بندی
☐	☐	☐	☐	☐	۳-۱. فازبندی زمان ساخت پیکره.....

پرسش‌نامه به همان ده متخصص موضوعی تحویل داده شد تا نظرات آن‌ها درباره اعمال یا رد مسائل مطرح شده در دور دوم دلفی، جمع‌آوری شود. جدول ۴-۱۳، درصد فراوانی انتخاب گزینه‌های کاملاً موافقم تا کاملاً مخالفم، را نشان می‌دهد.

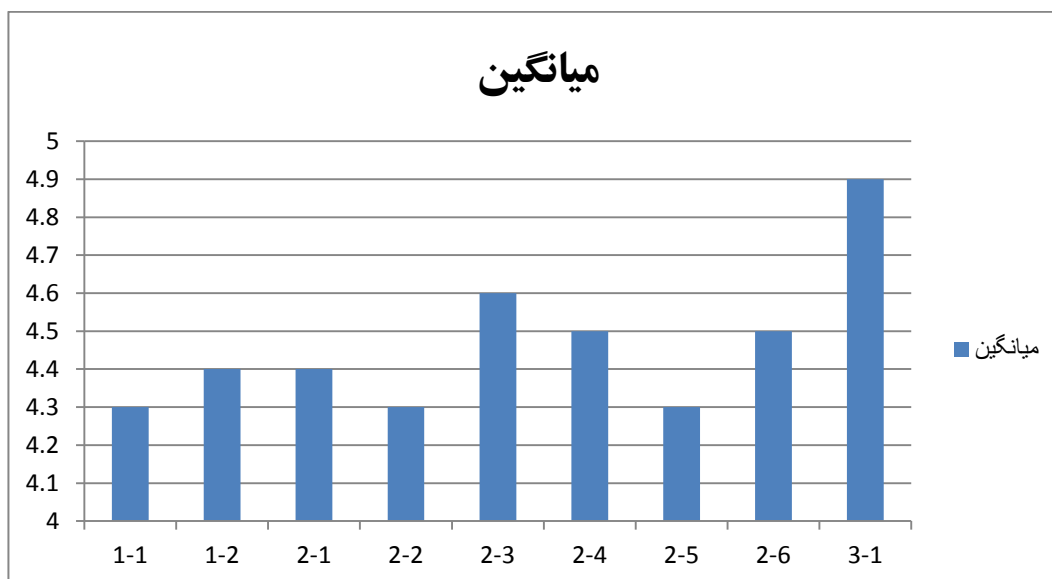
جدول ۴-۱۳: درصد فراوانی پاسخ به هر گویه در پرسش‌نامه دور دوم دلفی

شماره گویه	درصد فراوانی انتخاب گزینه ۵	درصد فراوانی انتخاب گزینه ۴	درصد فراوانی انتخاب گزینه ۳	درصد فراوانی انتخاب گزینه ۲	درصد فراوانی انتخاب گزینه ۱
	(کاملاً موافقم)	(تا اندازه‌ای موافقم)	(مطمئن نیستم)	(تا اندازه‌ای مخالفم)	(کاملاً مخالفم)
۱-۱	٪۶۰	٪۲۰	٪۱۰	٪۱۰	٪۰
۲-۱	٪۵۰	٪۴۰	٪۱۰	٪۰	٪۰
۱-۲	٪۶۰	٪۲۰	٪۲۰	٪۰	٪۰
۲-۲	٪۶۰	٪۱۰	٪۳۰	٪۰	٪۰
۳-۲	٪۷۰	٪۲۰	٪۱۰	٪۰	٪۰
۴-۲	٪۷۰	٪۱۰	٪۲۰	٪۰	٪۰
۵-۲	٪۶۰	٪۲۰	٪۱۰	٪۱۰	٪۰
۶-۲	٪۸۰	٪۰	٪۱۰	٪۱۰	٪۰
۱-۳	٪۹۰	٪۱۰	٪۰	٪۰	٪۰

جدول ۴-۱۴، و نمودار ۴-۵، میانگین پاسخ به هر گویه، در پرسش‌نامه دور دوم دلفی را نشان می‌دهد.

جدول ۴-۱۴: میانگین پاسخ به هر گویه در پرسش‌نامه دور دوم دلفی

شماره گویه	میانگین پاسخ‌ها
۱-۱	۴, ۳
۲-۱	۴, ۴
۱-۲	۴, ۴
۲-۲	۴, ۳
۳-۲	۴, ۶
۴-۲	۴, ۵
۵-۲	۴, ۳
۶-۲	۴, ۵
۷-۲	۴, ۹



نمودار ۴-۵: میانگین پاسخ به هر گویه در پرسش‌نامه دور دوم دلفی

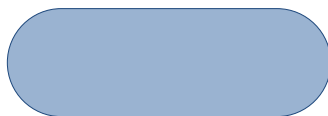
با توجه به جدول ۴-۱۴، می‌توان این‌طور نتیجه‌گیری نمود که چنانچه میانگین پاسخ به گویه‌ای ≤ 4 باشد، آن گویه برای امکان‌سنجی لحاظ می‌شود و گویه‌هایی که فاقد این شرایط باشند، از امکان‌سنجی حذف می‌شوند. همانگونه که در جدول ۴-۱۴ مشخص شده است، میانگین پاسخ به گویه‌ها بالاتر از ۴ است. بنابراین متخصصان با درج گویه‌ها در امکان‌سنجی اتفاق‌نظر دارند و همه گویه‌های موجود در این بخش باید در امکان‌سنجی لحاظ شوند.

۴-۶- مدل امکان‌سنجی ساخت پیکره زبانی (پیکره متنی)

با در نظر گرفتن نظرات متخصصان درباره مولفه‌های امکان‌سنجی ساخت پیکره، مدل امکان‌سنجی ساخت پیکره زبانی (نوعاً پیکره متنی) و ابعاد و مولفه‌های مربوطه در جدول ۴-۱۵ و شکل ۴-۶، قابل مشاهده است:

جدول ۴-۱۵: مدل امکان‌سنجی ساخت پیکره متنی

ابعاد	مؤلفه‌ها
۱. فنی	۱-۱. بررسی نیاز به زبان‌شناس همگانی/ عمومی در تیم ساخت پیکره
	۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره
	۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره
	۴-۱. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره
	۵-۱. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره
	۶-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره
	۷-۱. تعیین حجم متون، به عنوان گامی از ساخت پیکره
	۸-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره
	۹-۱. حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره
	۱۰-۱. بررسی نیاز به رابط کاربر (User interface) برای کار کردن کاربران با پیکره
	۱۱-۱. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند
۲. اقتصادی	۱-۲. برآورد هزینه متخصصان ساخت پیکره
	۲-۲. برآورد هزینه برنامه‌نویسی رایانه
	۳-۲. برآورد هزینه حاشیه‌نویسی دستی
	۴-۲. برآورد هزینه آموزش کاربران
	۵-۲. برآورد هزینه خرید نرم‌افزار
	۶-۲. برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و...)
	۷-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و...)
	۸-۲. برآورد هزینه نگهداری زیرساخت (سرور، شبکه، و...)
	۹-۲. برآورد هزینه ساخت ابزار گردآوری متن
	۱۰-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره
	۱۱-۲. تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)
	۱۲-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پایشگر/ بروزشونده
۳. زمان‌بندی	۱-۳. برآورد زمان‌بندی ساخت پیکره
	۲-۳. فازبندی زمان ساخت پیکره
۴. قانونی/ حقوقی	۱-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره
	۲-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره
	۳-۴. بررسی حقوق مادی پیکره (پس از ساخت)
۵. عملیاتی	۱-۵. بررسی پذیرش و آمادگی سازمان در کاربرد پیکره
	۲-۵. بررسی دستاوردهای سازمان با ساخت پیکره
۶. تأمین مالی	۱-۶. پیش‌بینی منابع تأمین مالی برای ساخت پیکره
۷. بازاریابی	۱-۷. شناخت کاربران پیکره
	۲-۷. آزمون عملکرد پیکره پیش از عرضه به بازار



شکل ۴-۶: مدل امکان‌سنجی ساخت پیکره متنی

همان‌گونه که این مدل نشان می‌دهد، هفت بعد و ۳۳ مولفه باید در امکان‌سنجی ساخت پیکره متنی در

نظر گرفته شود. ابعاد شامل: بعد فنی، بعد اقتصادی، بعد زمان‌بندی، بعد قانونی حقوقی، بعد عملیاتی، بعد تامین مالی و بعد بازاریابی است. در هر بعد مولفه‌هایی در نظر گرفته شده است که باید در ساخت پیکره متنی لحاظ شوند. در بعد فنی نیروی متخصص لازم در تیم ساخت پیکره شامل زبان‌شناس همگانی/عمومی، زبان‌شناس رایانشی، برنامه‌نویس رایانه‌ای، پژوهشگرانی که تجربه ساخت پیکره دارند، است، بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره و رابط کاربر (User interface) برای کار کردن کاربران با پیکره از دیگر ملاحظات فنی ساخت پیکره است. همچنین در بعد فنی گام‌هایی هم برای سخت پیکره در نظر گرفته شده است که شامل انتخاب نوع متون، تعیین حجم متون، پیش‌پردازش متون و حاشیه‌نویسی (برچسب‌گذاری و...) است. در بعد اقتصادی نیز باید برای مشاوره گرفتن و استفاده از نیروهای متخصص، برآورد هزینه صورت پذیرد. همچنین در مورد هزینه برنامه‌نویسی رایانه، حاشیه‌نویسی دستی، آموزش کاربران، خرید نرم‌افزار، خرید وسایل و تجهیزات اداری (رایانه، میز، صندلی و...)، تأمین زیرساخت (سرور، شبکه، و...)، نگهداری زیرساخت (سرور، شبکه، و...)، ساخت ابزار گردآوری متن، ساخت ابزار کار با داده‌های پیکره و بروز کردن داده‌های پیکره در پیکره‌های پایشگر/بروزشونده، نیز باید برآورد هزینه صورت پذیرد. از دیگر مواردی که در بعد اقتصادی لحاظ می‌شود، تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه) است. در بعد زمان‌بندی، باید زمان‌بندی پروژه ساخت پیکره و فازهای آن مشخص شود. مولفه‌های مورد بررسی در بعد حقوقی قانونی شامل الزام‌های قانونی و حقوقی، اجازه‌نامه کاربرد متون و حقوق مادی پیکره (پس از ساخت) است تا ساخت پیکره از نظر حقوقی قانونی با مانعی برخورد نکند. از دیگر ابعاد مورد بررسی در پروژه ساخت پیکره، تامین مالی است که باید منابع تامین مالی برای ساخت پیکره پیش‌بینی شود. در بعد بازاریابی نیز شناخت کاربران پیکره و آزمون عملکرد پیکره پیش از عرضه به بازار باید مورد بررسی قرار گیرد.

همچنین ابزارها و حاشیه‌نویسی‌های مورد استفاده در پیکره متنی می‌تواند شامل: ابزار جداسازی، ابزار نرمال‌سازی متن/ابزاری به منظور پیش‌پردازش متن، ابزار «پارسر»، ابزار ریشه‌یاب، ابزاری برای برچسب‌گذاری اجزای واژگانی کلام، ابزار تشخیص‌دهنده گروه‌های نحوی، و ابزار خطایاب، حاشیه‌نویسی نحوی، کاربردی، گفتمانی، سبکی، و واژگانی، ابزار حاشیه‌نویسی پیکره، ابزار شناسایی موجودیت‌های نامدار (named entity recognition)، خزش‌گر (crawler)، تعیین فراداده (موضوع متن، جنسیت نویسنده، سال انتشار و...)، حاشیه‌نویسی (نشانه‌گذاری) معنایی (semantic annotation)، در مورد حاشیه‌نویسی گفتمانی، مشخصاً حاشیه‌نویسی حرکتهای بلاغی (rhetorical moves) چکیده‌ها،

برچسب‌گذاری نقش‌های معنایی (semantic role labeling) و حاشیه‌نویسی استعاره (metaphor annotation)، باشد.

۴-۷- خلاصه فصل

در این فصل ابتدا به تعریف امکان‌سنجی و ابعاد گوناگون امکان‌سنجی پرداخته شد. سپس، اعتباریابی چارچوب مفهومی امکان‌سنجی ساخت پیکره توضیح داده شد که خود شامل نحوه تهیه پرسش‌نامه و تحویل آن به متخصصان مربوطه و توضیح دور اول و دوم دلفی است. در نهایت، پس از بررسی آماری نظرات متخصصان در دور اول و دوم دلفی، مدل امکان‌سنجی ساخت پیکره ارائه شد (شکل ۴-۶).

فصل پنجم

امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

۵-۱- مقدمه

ایرانداک مرکز ثبت و تنها آرشیو ملی اطلاعات پایان‌نامه‌ها، رساله‌ها، و پیشنهاد آن‌ها در حوزه‌های گوناگون فنی-مهندسی، علوم انسانی، علوم پایه و... است. افزون بر این، آرشیو پایان‌نامه‌ها و رساله‌های دانش‌آموختگان ایرانی خارج از کشور نیز در ایرانداک است و بیشترین طرح‌های پژوهشی و گزارش‌های دولتی در ایرانداک ثبت شده‌اند. بنابراین، ایرانداک منبع غنی داده‌های علمی منسجم و معتبر است. تهیه یک پیکره زبانی مبتنی بر داده‌های علمی ایرانداک از سه نظر حائز اهمیت است: ۱- ایرانداک را در شمار پژوهشگاه‌ها و دانشگاه‌هایی قرار می‌دهد که پیکره زبانی مفید تهیه کرده‌اند و پژوهشگران می‌توانند از این پیکره خاص، که پیکره‌ای مبتنی بر داده‌های علمی ایرانداک است، استفاده کنند، ۲- با توجه به حجم عظیم داده‌های علمی در ایرانداک (نوعاً پایان‌نامه‌ها)، با توجه به اینکه این اطلاعات علمی عظیم تنها در اختیار ایرانداک است، ساختن چنین پیکره‌ای کاملاً منحصر به فرد و بی‌رقیب است؛ تاکنون پیکره‌ای علمی مبتنی بر چنین داده‌های عظیمی ساخته نشده است و می‌تواند مورد استفاده جامعه دانشگاهی قرار گیرد و ۳- می‌توان بسیاری از پژوهش‌های حوزه پردازش متن (مانند انواع طبقه‌بندی داده‌ها، استخراج اطلاعات گوناگون، ریشه‌یابی کلمات، تقطیع کلمات، پردازش نحوی جملات و...) را با استفاده از این پیکره انجام داد. پیش از اقدام به ساخت پیکره متنی مبتنی بر داده‌های علمی ایرانداک، باید مراحل، ابزارهای مورد نیاز، تیم‌های تخصصی لازم، برآورد هزینه و... جهت ساخت پیکره‌ای مطلوب و بسامان مبتنی بر داده‌های ایرانداک مشخص شود تا بتوان درباره ساخت چنین پیکره‌ای در ایرانداک بهتر تصمیم‌گیری کرد و در صورت داشتن امکانات مالی و نیروی متخصص و ابزارهای لازم، در طرح / طرح‌های پژوهشی دیگر، اقدام به ساخت چنین پیکره‌ای کرد؛ مشخص نمودن این مراحل و هزینه‌ها و تیم‌های تخصصی لازم مستلزم انجام امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک است. در این مرحله از پژوهش، مدل امکان‌سنجی ساخت پیکره متنی طراحی شده است و باید به طراحی مدل امکان‌سنجی ساخت پیکره متنی در ایرانداک پرداخته شود. در این فصل، مدل مذکور طراحی خواهد شد و به اعتباریابی مدل پرداخته خواهد شد.

۵-۲- طراحی مدل امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

برای طراحی مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک باید نظرات متخصصان و کارشناسان حاضر در ایرانداک درباره ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره متنی، جمع‌آوری

گردد تا مشخص شود کدامیک از ابعاد و مولفه‌های مدل باید در طراحی مدل جدید برای ایرانداک لحاظ شوند و کدامیک حذف شوند و اگر ابعاد و مولفه/مولفه‌های دیگری نیز باید برای در طراحی مدل امکان-سنجی برای ایرانداک لحاظ شوند، در مدل وارد شوند. در این راستا پنلی تشکیل شد که شامل نه نفر متخصص علمی و اجرایی بود: رئیس پژوهشگاه، چهار نفر از اعضای هیات علمی گروه زبان‌شناسی رایانشی و اصطلاح‌شناسی و چهار نفر کارشناسان اجرایی شاغل در بخش‌های گوناگون ایرانداک بود. پیش از برگزاری پنل، ابتدا پرسش‌نامه‌ای تهیه شد که حاوی همان ابعاد و مولفه‌هایی است که در مدل کلی امکان‌سنجی ساخت پیکره متنی (فصل چهارم، جدول ۴-۱۵ و شکل ۴-۶) آورده شده است، این پرسش‌نامه در پنل در اختیار اعضای پنل قرار گرفت. پرسش‌نامه در جدول ۵-۱، قابل مشاهده است.

جدول ۵-۱: پرسش‌نامه ساخت پیکره متنی از داده‌های علمی ایرانداک

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین و نظر متخصصان به دست آمده‌اند.

با بودن هر یک از مولفه‌های زیر					در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
کاملاً موافقم ۵	تا اندازه-ای موافقم ۴	مطمئن نیستم ۳	تا اندازه-ای مخالفم ۲	کاملاً مخالفم ۱	
۱. بُعد فنی					
۱-۱. بررسی نیاز به زبان‌شناس همگانی/عمومی در تیم ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	
۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	
۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	
۴-۱. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	
۵-۱. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره					
☐ ۵	☐ ۴	☐ ۳	☐ ۲	☐ ۱	

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین و نظر متخصصان به دست آمده‌اند.

کاملاً مخالفم ۱	تا اندازه- ای مخالفم ۲	مطمئن نیستم ۳	تا اندازه- ای موافقم ۴	کاملاً موافقم ۵	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
☐	☐	☐	☐	☐	۶-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۷-۱. تعیین حجم متون، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۸-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۹-۱. حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره.....
☐	☐	☐	☐	☐	۱۰-۱. بررسی نیاز به رابط کاربر (User interface) برای کار کردن کاربران با پیکره.....
☐	☐	☐	☐	☐	۱۱-۱. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند.....
.....					
۲- بُعد اقتصادی					
☐	☐	☐	☐	☐	۱-۲. برآورد هزینه متخصصان ساخت پیکره.....
☐	☐	☐	☐	☐	۲-۲. برآورد هزینه برنامه‌نویسی رایانه.....

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین و نظر متخصصان به دست آمده‌اند.

کاملاً موافقم ۵	تا اندازه-ای موافقم ۴	مطمئن نیستم ۳	تا اندازه-ای مخالفم ۲	کاملاً مخالفم ۱	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
☐	☐	☐	☐	☐	۳-۲. برآورد هزینه حاشیه‌نویسی دستی
☐	☐	☐	☐	☐	۴-۲. برآورد هزینه آموزش کاربران
☐	☐	☐	☐	☐	۵-۲. برآورد هزینه خرید نرم‌افزار
☐	☐	☐	☐	☐	۶-۲. برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و.....)
☐	☐	☐	☐	☐	۷-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و.....)
☐	☐	☐	☐	☐	۸-۲. برآورد هزینه نگهداری زیرساخت (سرور، شبکه، و.....)
☐	☐	☐	☐	☐	۹-۲. برآورد هزینه ساخت ابزار گردآوری متن
☐	☐	☐	☐	☐	۱۰-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره
☐	☐	☐	☐	☐	۱۱-۲. تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)
☐	☐	☐	☐	☐	۱۲-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پایشگر / پژوهشونده

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین و نظر متخصصان به دست آمده‌اند.

کاملاً موافقم ۵	تا اندازه- ای موافقم ۴	مطمئن نیستم ۳	تا اندازه- ای مخالفم ۲	کاملاً مخالفم ۱	با بودن هر یک از مولفه‌های زیر در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟
.....					

۳- بُعد زمان‌بندی

۳-۱. برآورد زمان‌بندی ساخت پیکره □۵ □۴ □۳ □۲ □۱

.....

۳-۲. فازبندی زمان ساخت پیکره □۵ □۴ □۳ □۲ □۱

.....

۴- بُعد قانونی / حقوقی

۴-۱. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره □۵ □۴ □۳ □۲ □۱

.....

۴-۲. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره □۵ □۴ □۳ □۲ □۱

.....

۴-۳. بررسی حقوق مادی پیکره (پس از ساخت) □۵ □۴ □۳ □۲ □۱

.....

۵- بُعد عملیاتی

۵-۱. بررسی پذیرش و آمادگی سازمان در کاربرد پیکره □۵ □۴ □۳ □۲ □۱

.....

۵-۲. بررسی دستاوردهای سازمان با ساخت پیکره □۵ □۴ □۳ □۲ □۱

.....

دیدگاه شما درباره مؤلفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی پژوهش‌های پیشین و نظر متخصصان به دست آمده‌اند.

با بودن هر یک از مولفه‌های زیر				
کاملاً موافقم ۵	تا اندازه-ای موافقم ۴	مطمئن نیستم ۳	تا اندازه-ای مخالفم ۲	کاملاً مخالفم ۱
در امکان‌سنجی ساخت پیکره متنی چه اندازه موافق هستید؟				

۶- بُعد تأمین منابع

۱-۶. پیش‌بینی منابع تأمین مالی برای ساخت پیکره □۵ □۴ □۳ □۲ □۱

۷- بُعد بازاریابی

۱-۷. شناخت کاربران پیکره □۵ □۴ □۳ □۲ □۱

۲-۷. آزمون عملکرد پیکره پیش از عرضه به بازار □۵ □۴ □۳ □۲ □۱

توضیحات مربوط به هر گویه به پاسخ‌گویان به طور شفاهی ارائه شد و هر جا سوالی مطرح شد، به سوالات پاسخ داده شد تا هیچ گونه ابهامی درباره پرسش‌نامه باقی نماند.

۵-۳- اعتباریابی مدل امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

پس از جمع‌آوری پرسش‌نامه ارائه شده در پیل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک، پاسخ‌ها به دقت مورد بررسی قرار گرفت. جدول ۵-۲، درصد فراوانی انتخاب گزینه‌های کاملاً موافقم تا کاملاً مخالفم، را نشان می‌دهد.

جدول ۵-۲: درصد فراوانی پاسخ به هر گویه در پرسش‌نامه ساخت پیکره متنی از داده‌های علمی ایرانداک

شماره گویه	درصد فراوانی انتخاب گزینه ۵	درصد فراوانی انتخاب گزینه ۴	درصد فراوانی انتخاب گزینه ۳	درصد فراوانی انتخاب گزینه ۲	درصد فراوانی انتخاب گزینه ۱
۱-۱	٪۷۰	٪۳۰	٪۰	٪۰	٪۰
۲-۱	٪۱۰۰	٪۰	٪۰	٪۰	٪۰
۳-۱	٪۸۰	٪۲۰	٪۰	٪۰	٪۰
۴-۱	٪۵۰	٪۰	٪۱۰	٪۳۰	٪۱۰
۵-۱	٪۱۰۰	٪۰	٪۰	٪۰	٪۰
۶-۱	٪۷۰	٪۲۰	٪۱۰	٪۰	٪۰
۷-۱	٪۷۰	٪۳۰	٪۰	٪۰	٪۰
۸-۱	٪۸۰	٪۱۰	٪۱۰	٪۰	٪۰
۹-۱	٪۹۰	٪۱۰	٪۰	٪۰	٪۰
۱۰-۱	٪۵۰	٪۳۰	٪۲۰	٪۰	٪۰
۱۱-۱	٪۶۰	٪۱۰	٪۳۰	٪۰	٪۰
۱-۲	٪۹۰	٪۱۰	٪۰	٪۰	٪۰
۲-۲	٪۹۰	٪۱۰	٪۰	٪۰	٪۰
۳-۲	٪۸۰	٪۱۰	٪۱۰	٪۰	٪۰
۴-۲	٪۳۰	٪۳۰	٪۰	٪۳۰	٪۱۰
۵-۲	٪۷۰	٪۱۰	٪۰	٪۲۰	٪۰
۶-۲	٪۳۰	٪۳۰	٪۱۰	٪۲۰	٪۰
۷-۲	٪۵۰	٪۳۰	٪۱۰	٪۱۰	٪۰
۸-۲	٪۵۰	٪۰	٪۱۰	٪۳۰	٪۱۰
۹-۲	٪۳۰	٪۱۰	٪۰	٪۴۰	٪۲۰
۱۰-۲	٪۵۰	٪۳۰	٪۰	٪۲۰	٪۰
۱۱-۲	٪۴۰	٪۳۰	٪۱۰	٪۲۰	٪۰
۱۲-۲	٪۶۰	٪۳۰	٪۱۰	٪۰	٪۰
۱-۳	٪۹۰	٪۱۰	٪۰	٪۰	٪۰
۲-۳	٪۸۰	٪۲۰	٪۰	٪۰	٪۰
۱-۴	٪۷۰	٪۳۰	٪۰	٪۰	٪۰
۲-۴	٪۵۰	٪۳۰	٪۰	٪۲۰	٪۰

امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۱۲۰

۳-۴	%۳۰	%۳۰	%۰	%۱۰	%۳۰
۱-۵	%۲۰	%۱۰	%۲۰	%۳۰	%۲۰
۲-۵	%۵۰	%۰	%۳۰	%۰	%۲۰
۱-۶	%۸۰	%۱۰	%۰	%۰	%۱۰
۱-۷	%۷۰	%۲۰	%۱۰	%۰	%۰
۲-۷	%۷۰	%۰	%۲۰	%۰	%۱۰

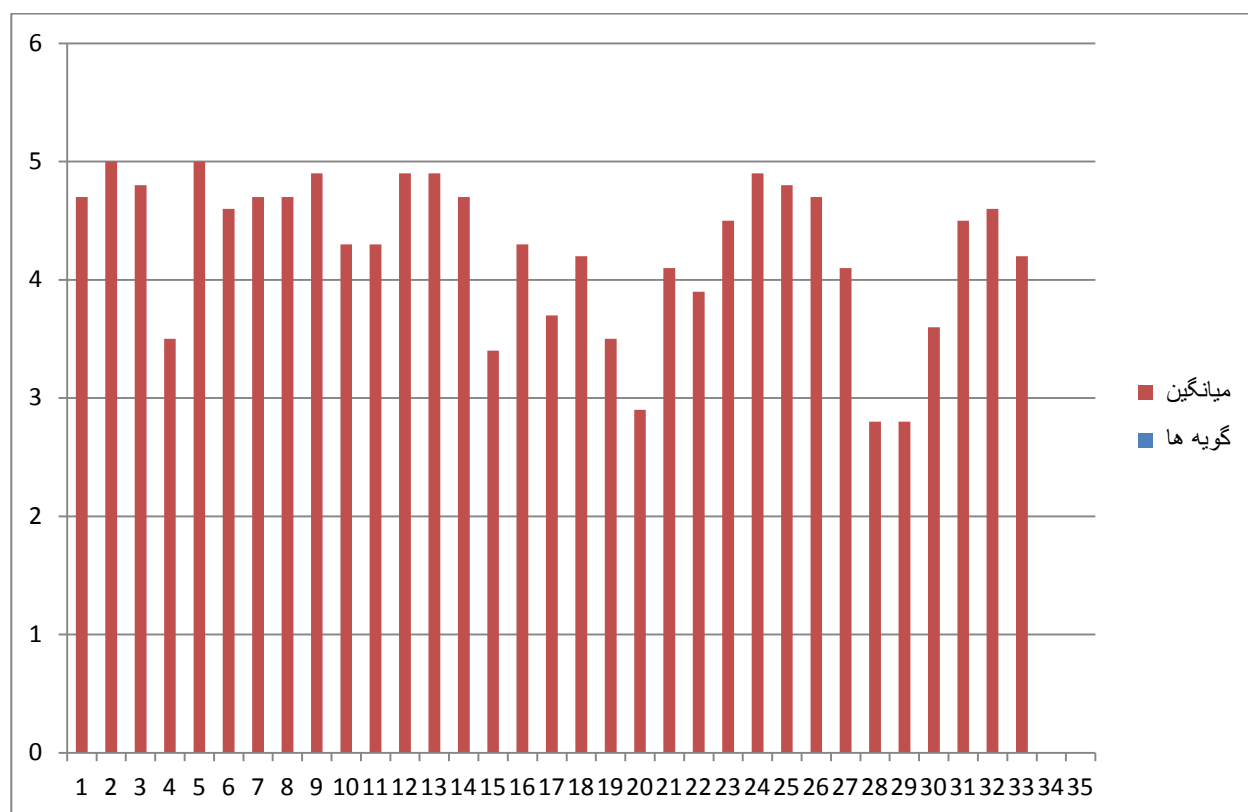
جدول ۳-۵ و نمودار ۱-۵، میانگین پاسخ به هر گویه در امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک را نشان می‌دهد.

جدول ۳-۵: میانگین پاسخ به هر گویه در امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

شماره گویه	میانگین پاسخ به هر گویه
۱-۱	۴,۷
۲-۱	۵
۳-۱	۴,۸
۴-۱	۳,۵
۵-۱	۵
۶-۱	۴,۶
۷-۱	۴,۷
۸-۱	۴,۷
۹-۱	۴,۹
۱۰-۱	۴,۳
۱۱-۱	۴,۳
۱-۲	۴,۹
۲-۲	۴,۹
۳-۲	۴,۷
۴-۲	۳,۴
۵-۲	۴,۳
۶-۲	۳,۷
۷-۲	۴,۲
۸-۲	۳,۵
۹-۲	۲,۹
۱۰-۲	۴,۱
۱۱-۲	۳,۹

امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۱۲۱

۱۲-۲	۴,۵
۱-۳	۴,۹
۲-۳	۴,۸
۱-۴	۴,۷
۲-۴	۴,۱
۳-۴	۲,۸
۱-۵	۲,۸
۲-۵	۳,۶
۱-۶	۴,۵
۱-۷	۴,۶
۲-۷	۴,۲



نمودار ۱-۵: میانگین پاسخ به هر گوپه در امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

با توجه به جدول ۵-۳، می‌توان اینطور نتیجه‌گیری نمود که چنانچه میانگین پاسخ به گویه‌ای ۴ ≤ باشد، آن گویه برای امکان‌سنجی لحاظ می‌شود و گویه‌هایی که فاقد این شرایط باشند، از امکان‌سنجی حذف می‌شوند. همانگونه که در جدول ۵-۳ مشخص شده است، میانگین پاسخ به همه گویه‌ها، غیر از گویه‌های ۴-۱، ۴-۲، ۴-۵، ۶-۲، ۸-۲، ۹-۲، ۱۱-۲، ۳-۴، ۵-۱ و ۵-۲، بالاتر از ۴ است. بنابراین اکثر متخصصان با درج گویه‌ها در امکان‌سنجی اتفاق نظر دارند و همه گویه‌های موجود در بخش نخست (غیر از گویه‌های ۴-۱، ۴-۲، ۴-۵، ۶-۲، ۸-۲، ۹-۲، ۱۱-۲، ۳-۴، ۵-۱ و ۵-۲) باید در امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک لحاظ شوند. گویه‌های حذف شده شامل موارد زیر هستند:

گویه ۴-۱: بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره

گویه ۴-۲: برآورد هزینه آموزش کاربران

گویه ۶-۲: برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و...)

گویه ۸-۲: برآورد هزینه نگهداری زیرساخت (سرور، شبکه، و...)

گویه ۹-۲: برآورد هزینه ساخت ابزار گردآوری متن

گویه ۱۱-۲: تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)

گویه ۳-۴: بررسی حقوق مادی پیکره (پس از ساخت)

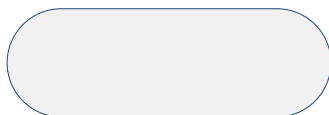
گویه ۱-۵: بررسی پذیرش و آمادگی سازمان در کاربرد پیکره

گویه ۲-۵: بررسی دستاوردهای سازمان با ساخت پیکره

بنابراین، یک گویه از بعد فنی، پنج گویه از بعد اقتصادی، یک گویه از بعد قانونی/حقوقی حذف شده‌اند و بعد عملیاتی هم به کل از مدل امکان‌سنجی ساخت پیکره متنی حذف شده است. با در نظر گرفتن گویه‌های باقیمانده، مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک در قالب جدول ۵-۴ و نمودار ۵-۲، قابل مشاهده است:

جدول ۴-۵: مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک

ابعاد	مولفه‌ها
۱. فنی	۱-۱. بررسی نیاز به زبان‌شناس همگانی/ عمومی در تیم ساخت پیکره ۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره ۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره ۴-۱. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره ۵-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره ۶-۱. تعیین حجم متون، به عنوان گامی از ساخت پیکره ۷-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره ۸-۱. حاشیه‌نویسی (برجسب‌گذاری و...)، به عنوان گامی از ساخت پیکره ۹-۱. بررسی نیاز به رابط کاربر (User interface) برای کار کردن کاربران با پیکره ۱۰-۱. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند
۲. اقتصادی	۱-۲. برآورد هزینه متخصصان ساخت پیکره ۲-۲. برآورد هزینه برنامه‌نویسی رایانه ۳-۲. برآورد هزینه حاشیه‌نویسی دستی ۴-۲. برآورد هزینه خرید نرم‌افزار ۵-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و...) ۶-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره ۷-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پیشگر/ بروزشونده
۳. زمان‌بندی	۱-۳. برآورد زمان‌بندی ساخت پیکره ۲-۳. فازبندی زمان ساخت پیکره
۴. قانونی / حقوقی	۱-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره ۲-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره
۵. تامین مالی	۱-۵. پیش‌بینی منابع تأمین مالی برای ساخت پیکره
۶. بازاریابی	۱-۶. شناخت کاربران پیکره ۲-۶. آزمون عملکرد پیکره پیش از عرضه به بازار



شکل ۵-۲: مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک

همان‌گونه که این مدل نشان می‌دهد، شش بعد و ۲۴ مولفه باید در امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک در نظر گرفته شود. ابعاد شامل: بعد فنی، بعد اقتصادی، بعد زمان‌بندی، بعد قانونی حقوقی، بعد تامین مالی و بعد بازاریابی است. در هر بعد مولفه‌هایی در نظر گرفته شده است که باید در ساخت پیکره متنی از داده‌های علمی ایرانداک لحاظ شوند. در بعد فنی ده مولفه شامل: نیروی

متخصص لازم در تیم ساخت پیکره که خود شامل زبان‌شناس همگانی / عمومی، زبان‌شناس رایانشی، برنامه‌نویس رایانه‌ای، پژوهشگرانی که تجربه ساخت پیکره دارند، است، رابط کاربر (User interface) برای کار کردن کاربران با پیکره، گام‌های ساخت پیکره که شامل طراحی زیربنای پیکره، انتخاب نوع متون، تعیین حجم متون، پیش‌پردازش متون و حاشیه‌نویسی (برچسب‌گذاری و...) است. در بعد اقتصادی نیز هفت مولفه باید در نظر گرفته شود که شامل: برآورد هزینه مشاوره گرفتن و استفاده از نیروهای متخصص، هزینه برنامه‌نویسی رایانه، حاشیه‌نویسی دستی، خرید نرم‌افزار، تأمین زیرساخت (سرور، شبکه، و...)، ساخت ابزار کار با داده‌های پیکره و بروز کردن داده‌های پیکره در پیکره‌های پایشگر / بروزشونده، است. در بعد زمان‌بندی، باید زمان‌بندی پروژه ساخت پیکره و فازهای آن مشخص شود. مولفه‌های مورد بررسی در بعد حقوقی قانونی شامل الزام‌های قانونی و حقوقی و اجازه‌نامه کاربرد متون است تا ساخت پیکره از نظر حقوقی قانونی با مانعی برخورد نکند. از دیگر ابعاد مورد بررسی در پروژه ساخت پیکره، تأمین مالی است که باید منابع تأمین مالی برای ساخت پیکره پیش‌بینی شود. در بعد بازاریابی نیز شناخت کاربران پیکره و آزمون عملکرد پیکره پیش از عرضه به بازار باید مورد بررسی قرار گیرد.

۵-۴- امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک

یکی از مسائل مهم در امکان‌سنجی انجام یک پروژه، ارزیابی اقتصادی آن است که صرفه اقتصادی انجام پروژه از پیش تعیین شود. به منظور دستیابی به حداکثر سود ممکن در سرمایه‌گذاری، باید پیش از هرگونه تصمیم‌گیری در مورد ورود به پروژه، مطالعات دقیق و جامع از جنبه‌های مختلف بر روی پروژه انجام گیرد. این کار با انجام امکان‌سنجی انجام پروژه (ساخت پیکره زبانی از داده‌های علمی ایرانداک) محیا شده است. حال با در دست داشتن مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک، می‌توان به بررسی این مسئله پرداخت که در نظر گرفتن ابعاد و مولفه‌های امکان‌سنجی چه هزینه‌هایی برای ایرانداک دارد و در نتیجه، به این وسیله بهتر بتوان درباره انجام چنین پروژه‌ای در ایرانداک بهتر تصمیم‌گیری نمود. یکی از روش‌هایی که بوسیله آن می‌توان ارزیابی اقتصادی انجام پروژه را انجام داد، روش تحلیل هزینه-فایده^۱ است. تحلیل هزینه-فایده، روشی برای ارزیابی مزیت نسبی پروژه‌های سرمایه‌گذاری بر حسب تخصیص بهینه و کارآمد منابع است. هدف تحلیل هزینه-فایده، بهبود کارایی منابع در جهت رفاه اقتصادی است. به عبارتی دیگر، هدف از ارزیابی، کمک کردن به انتخاب بهترین نوع تصمیم‌گیری

¹ Cost-benefit analysis

در جهت استفاده بهینه و مطلوب از منابع است (ویک: ۱۹۹۳، نقل از لطفعلی‌پور و اسلامی گیسکی: ۱۳۸۶). در راستای استفاده از این روش، ابتدا باید هزینه‌های مربوط به ساخت پیکره متنی از داده‌های ایرانداک مشخص شود. با مشاوره گرفتن از متخصصان موضوعی و کسانی که در زمینه ساخت پیکره، تجربه دارند، هزینه‌های مولفه‌های لحاظ شده در مدل امکان‌سنجی ساخت پیکره متنی در جدول ۵-۵، قابل مشاهده است. در این جدول برخی از ابعاد و مولفه‌ها که باید برآورد هزینه در موردشان انجام می‌گرفت، (مانند ابعاد فنی و اقتصادی) هزینه مربوطه محاسبه شده است و در جدول قابل مشاهده است. ابعاد دیگر مانند زمان‌بندی، بازاریابی، حقوقی/قانونی و تامین منابع مالی نیز چگونگی تصمیم‌گیری درباره وضعیت آن‌ها در جدول قابل مشاهده است. در این جدول، منظور از هزینه همکاری متخصصان و مشاوره با آن‌ها، مشاوره و همکاری در زمینه چگونگی طراحی زیربنای پیکره، انتخاب متون، تعیین حجم متون، نحوه پیش‌پردازش متون، تعیین نوع حاشیه‌نویسی و بررسی امکان استفاده از ابزارهای مربوطه یا انجام دستی حاشیه‌نویسی است. فقط باید در نظر داشته باشیم که تعیین هزینه مولفه‌ها، امری است غیرعینی / شخصی^۱، بنابراین، هنگام انجام پروژه ساخت پیکره، با توجه به این مسئله که تیم اجرایی برای هزینه‌ها چه میزانی را پیشنهاد کند، ممکن است تغییرات اجتناب‌ناپذیر و غیرقابل پیش‌بینی در میزان این هزینه‌ها صورت پذیرد. به هر حال، در پژوهش حاضر سعی شده است از متخصصان این حوزه که اطلاعات و تجربه کافی دارند، نظرخواهی شود.

در مدل کلی امکان‌سنجی ساخت پیکره متنی، مولفه «تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)» وجود داشت که در مدل امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک، میانگین کمتر از ۴ آورد و بنابراین حذف شد. اما از آنجایی که این مولفه در تصمیم‌گیری‌های اقتصادی و پیش‌گیری از رفع ابهام اهمیت دارد، در بعد اقتصادی در نظر گرفته شده است تا مشخص شود، ایرانداک در این زمینه چه تصمیمی خواهد گرفت. چنانچه تصمیم به استفاده رایگان باشد، دیگر نیازی به انجام مرحله «تعیین سود/فایده» اقتصادی انجام پروژه، نخواهد بود. اما چنانچه قرار نباشد پیکره به رایگان در اختیار کاربران قرار گیرد، می‌توان با در نظر گرفتن سناریوهای احتمالی به محاسبه درآمد حاصل از استفاده سالانه کاربران یا فروش آن پرداخت. هزینه‌های ثابت و متغیر و همچنین چگونگی پرداختن به مولفه‌های غیرمالی مدل امکان‌سنجی، در جدول ۵-۵، قابل مشاهده است.

¹ subjective

جدول ۵-۵: مشخص نمودن چگونگی پرداختن به هر یک از ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک

ابعاد امکان‌سنجی	تصمیم‌گیری یا مشخص نمودن هزینه مربوطه (به ریال)	نوع هزینه (ثابت یا متغیر)	جمع هزینه‌های ثابت به ریال	جمع هزینه‌های متغیر به ریال (سالانه)
۱- بعد فنی			700,000,000 (هفتاد میلیون تومان)	560,000,000 (پنجاه و شش میلیون تومان)
۱-۱. مشاوره با زبان‌شناس همگانی/عمومی	50,000,000 (پنج میلیون تومان)	ثابت	هزینه ثابت شامل موارد زیر است:	هزینه متغیر شامل موارد زیر است:
۱-۲. مشاوره با زبان‌شناس رایانشی	100,000,000 (ده میلیون تومان)	ثابت	۱. مشاوره با زبان‌شناس همگانی/عمومی (در فاز اول و دوم)	۱. هزینه تامین و نگهداری زیرساخت (سرور، شبکه و....)
۱-۳. همکاری برنامه‌نویس رایانه‌ای	500,000,000 (پنجاه میلیون تومان)	برای فاز اول: ثابت (وظایف ایشان شامل: طراحی زیربنای پیکره (پانزده میلیون تومان)، پیش‌پردازش متون (ده میلیون تومان)، طراحی رابط کاربر و ساخت ابزار کار با داده‌های پیکره (بیست میلیون) است) و پشتیبانی نرم-افزاری (برنامه‌نویسی) که البته هزینه آن متغیر است و سالانه محاسبه می‌گردد.	۲. مشاوره با زبان‌شناس رایانشی که تجربه ساخت پیکره هم داشته باشند (در فاز اول و دوم)	۲. هزینه بروز کردن داده‌های پیکره در پیکره-های پایشگر/بروزشوند ه
۱-۴. همکاری پژوهشگرانی	اگر از متخصصان زبان‌شناسی رایانشی که تجربه ساخت پیکره دارند، استفاده	ثابت	۳. همکاری برنامه‌نویس رایانه‌ای وظایف زیر را بر عهده دارد: الف. طراحی زیربنای پیکره ب. پیش‌پردازش متون پ. طراحی رابط کاربر و ساخت ابزار کار با داده‌های پیکره (مبلغ پنج میلیون تومان هم برای مسائل پیش-بینی نشده که ممکن است هنگام کار به وظایف برنامه‌نویس	۳. پشتیبانی نرم-افزاری / برنامه-نویسی

	اضافه شود، در نظر گرفته شده است) ۴. خرید نرم‌افزار		شود، هزینه آن‌ها در بخش «مشاوره با زبان‌شناس رایانشی» دیده شده است و دیگر نیازی نیست در این قسمت وارد شود. اما اگر زبان‌شناس/زبان‌شناسان رایانشی حاضر در تیم، تجربه ساخت پیکره نداشته باشند، می‌توان هزینه‌ای حدود 100,000,000 (ده میلیون تومان) را برای مشاوره با پژوهشگرانی که تجربه ساخت پیکره دارند، در نظر گرفت.	که تجربه ساخت پیکره دارند
		ثابت	150,000,000 (پانزده میلیون تومان) (این هزینه در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	۵-۱. طراحی زیربنای پیکره (چارچوب و ساختار)
			با مشاوره متخصصان مشخص خواهد شد	۶-۱. انتخاب نوع متون
			با مشاوره متخصصان مشخص خواهد شد	۷-۱. تعیین حجم متون
		ثابت	100,000,000 (ده میلیون تومان) (این هزینه در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	۸-۱. پیش‌پردازش متون
			به حجم داده‌های پیکره بستگی دارد و باید در فاز دوم که حجم پیکره مشخص شده است، این هزینه محاسبه شود.	۹-۱. حاشیه‌نویسی (برچسب-گذاری و...)
		ثابت	150,000,000 (پانزده میلیون تومان) (این بخش در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	۱۰-۱. رابط کاربر (user interface)

۲- بعد اقتصادی			
			۱-۲. هزینه همکاری متخصصان در بعد فنی دیده شده است
			۲-۲. هزینه برنامه-نویس رایانه-ای در بعد فنی دیده شده است
			۳-۲. هزینه حاشیه-نویسی دستی به حجم داده‌های پیکره بستگی دارد
		ثابت	۴-۲. هزینه خرید نرم‌افزار 50,000,000 (پنج میلیون تومان)
		متغیر (سالانه)	۵-۲. هزینه تامین و نگهداری زیرساخت (سرور، شبکه و....) 360,000,000 (ماهانه سه میلیون تومان که سالانه سی و شش میلیون تومان، می‌شود؛ تامین و نگهداری زیرساخت به بخش فنی پژوهشگاه مربوط می‌شود)
		ثابت	۶-۲. هزینه ساخت ابزار کار با داده‌های پیکره 50,000,000 (پنج میلیون تومان) (این بخش در واقع بخشی از کار طراحی رابط کاربر است، اما چون در مدل امکان‌سنجی به عنوان بخشی جدا در نظر گرفته شده، هزینه هم جدا آورده شده است)
		متغیر (سالانه)	۷-۲. هزینه بروز کردن داده-های پیکره در پیکره‌های پایشگر/بروزشونده 100,000,000 (ده میلیون تومان)
			۸-۲. تصمیم‌گیری درباره چگونگی دسترسی ایرانداک تصمیم دارد این پیکره را به رایگان در اختیار کاربران قرار دهد.

				کاربران به پیکره (رایگان یا با پرداخت هزینه)
		متغیر (سالانه)	بخش پشتیبانی شامل دو قسمت است: ۱. پشتیبانی از مسائل کاربران: در این بخش باید کسی در پژوهشگاه همکاری کند تا اگر نیازی به پاسخ‌گویی به درخواست یا ایمیل کاربران باشد، این کار را انجام دهد. بنابراین، پژوهشگاه می‌تواند پس از ساخت پیکره، درباره این بخش، تصمیم‌گیری کند. ۲. پشتیبانی نرم‌افزاری / برنامه‌نویسی: اگر نیازهای جدیدی از طرف کاربران مطرح گردد و نیاز به ایجاد تغییر یا اضافه کردن feature هایی به پیکره باشد، برنامه‌نویس رایانه‌ای تیم ساخت پیکره می‌تواند این کارها را انجام دهد که هزینه سالانه‌ای می‌توان برای آن در نظر گرفت. 100,000,000 (ده میلیون تومان)	۹-۲. پشتیبانی
				۳- بعد زمان‌بندی
			می‌توان پیکره را در دو فاز دید که طی یک الی دو سال پیکره کلی ساخته شود.	۱-۳. برآورد زمان‌بندی ساخت پیکره
		فاز اول: هزینه ثابت (هفتاد میلیون تومان) فاز دوم: مشخص نیست و به نوع حاشیه‌نویسی و	فاز اول: تشکیل تیم ساخت پیکره، طراحی زیربنای پیکره، پیش‌پردازش متون و وارد کردن متون در پیکره فاز دوم:	۲-۳. فازبندی زمان ساخت پیکره

تکات مهم جدول ۵-۵:

- پروژه ساخت پیکره از داده‌های علمی ایرانداک، می‌تواند در دو فاز تعریف شود که هزینه‌های فاز نخست، ثابت است و هزینه‌های فاز دوم بستگی به حجم پیکره و نوع حاشیه‌نویسی دارد (که هر دو این موارد در فاز نخست مشخص خواهند شد).
- در پایان فاز اول، در واقع پیکره ساخته شده است، اما پیکره‌ای است حاشیه‌نویسی نشده و برچسب‌گذاری داده‌ها انجام نشده است. تا همین مرحله هم می‌توان پایان پروژه ساخت پیکره را اعلام کرد، زیرا از پیکره حاشیه‌نویسی نشده هم می‌توان استفاده کرد. به عنوان نمونه، از این نوع پیکره‌ها می‌توان برای انواع طبقه‌بندی^۱ اطلاعات استفاده کرد. البته می‌توان در همین مرحله برخی از انواع حاشیه‌نویسی‌های اتوماتیک را هم وارد کرد. اما برای بالا بردن کارایی پیکره، در فاز اول تصمیم می‌گیریم کارایی پیکره را با انواع حاشیه‌نویسی بالا ببریم که در آن صورت باید حجم متون و نوع حاشیه‌نویسی معلوم شده باشد تا بتوان آن نوع حاشیه‌نویسی را ساخت یا دستی وارد و کنترل کرد.
- در فاز اول تیم ساخت پیکره تشکیل می‌شود. این تیم متشکل از زبان‌شناس عمومی، زبان‌شناس رایانشی، برنامه‌نویس رایانه‌ای و پژوهشگر/پژوهشگرانی است که تجربه ساخت پیکره دارند. در این پروژه می‌توان از برنامه‌نویس رایانه‌ای دعوت به همکاری شود که حرفه‌ای باشد و خود توانایی طراحی و ساخت نرم‌افزارهای مورد نیاز را داشته باشد. بنابراین بیشترین هزینه برای برنامه‌نویس رایانه‌ای در نظر گرفته شده است. کارهایی که برنامه‌نویس رایانه‌ای انجام خواهد داد شامل: طراحی زیربنای پیکره، پیش‌پردازش متون، طراحی رابط کاربر و ساخت ابزار کار با داده‌های پیکره، وارد کردن متون و پشتیبانی نرم‌افزاری/برنامه‌نویسی است (هزینه همه کارها به غیر از پشتیبانی، ثابت است). جزئیات هزینه‌های مربوط به متخصصان دیگر در جدول آورده شده است. همانگونه که در جدول مشخص شده است، اگر در تیم ساخت پیکره از زبان‌شناسان رایانشی دعوت به همکاری شود که تجربه ساخت پیکره هم دارند، دیگر نیازی نیست هزینه‌ای جدا برای همکاری پژوهشگرانی که تجربه ساخت پیکره دارند، در نظر گرفته شود. بنابراین، گویی دو کار با هم انجام می‌شود و در هزینه‌ها کمی صرفه‌جویی می‌شود.

¹ clustering

➤ هزینه حاشیه‌نویسی در جدول وارد نشده است. زیرا بستگی به حجم متون و نوع حاشیه‌نویسی دارد که در فاز اول درباره این موارد تصمیم‌گیری خواهد شد.

➤ هزینه‌های متغیر در بعد اقتصادی قابل مشاهده است. این هزینه‌ها شامل: هزینه تامین و نگهداری زیرساخت (سرور، شبکه و....)، هزینه بروز کردن داده‌های پیکره در پیکره‌های پیشگرا/بروزشونده و هزینه پشتیبانی است. بخش پشتیبانی خود شامل دو قسمت است: الف. پشتیبانی از مسائل کاربران: در این بخش باید کسی در سازمان همکاری کند تا اگر نیازی به پاسخ‌گویی به درخواست یا ایمیل کاربران باشد، این کار را انجام دهد. بنابراین، پژوهشگاه می‌تواند پس از ساخت پیکره، درباره این بخش، تصمیم‌گیری کند. ب. پشتیبانی نرم‌افزاری/ برنامه‌نویسی: اگر نیازهای جدیدی از طرف کاربران مطرح گردد و نیاز به ایجاد تغییر یا اضافه کردن feature هایی به پیکره باشد، برنامه‌نویس رایانه‌ای تیم ساخت پیکره می‌تواند این کارها را انجام دهد که هزینه سالانه‌ای می‌توان برای آن در نظر گرفت.

ابعاد حقوقی پروژه ساخت پیکره: در سایت‌های داخلی چیزی درباره دریافت اجازه‌نامه برای ساخت پیکره موجود نبود. در سایت‌های خارجی، مسائل متعددی در این حوزه مطرح شده است. به عنوان نمونه، در یکی از منابع غیرفارسی‌زبان، درباره دریافت اجازه‌نامه اینگونه بیان شده است: درجه ضرورت دریافت اجازه‌نامه برای ساخت پیکره، به صورت طیف است. به عنوان نمونه، دو شرط برای عدم نیاز به اجازه‌نامه برای ساخت پیکره آورده شده است (<https://translationjournal.net/journal/36corpus.htm>):

الف. پیکره حاوی متونی باشد که به صورت رایگان در دسترس عموم باشد.

ب. پیکره برای اهداف غیرتجاری استفاده شود.

برای اطمینان از قرار گرفتن در مسیر درست حقوقی در راستای ساخت پیکره از داده‌های علمی ایرانداک، نگارنده با سه مشاور حقوقی (آقای دکتر صفر بیگ‌زاده، آقای دکتر رحیم پیلوار و آقای سیامک قاجار) مصاحبه کرده است. آقای دکتر پیلوار مشخصاً به ماده ۵ و ماده ۸ قانون حمایت از حقوق مولفان، مصنفان و هنرمندان اشاره کردند. شرح این دو ماده به این صورت است:

ماده ۵ - پدیدآورنده اثرهای مورد حمایت این قانون می‌تواند استفاده از حقوق مادی خود را در کلیه موارد و از جمله موارد زیر به غیر واگذار نماید.

- ۱- تهیه فیلمهای سینمایی و تلویزیونی و مانند آن
- ۲- نمایش صحنه‌ای مانند تئاتر و باله و نمایشهای دیگر
- ۳- ضبط تصویری یا صوتی اثر بر روی صفحه یا نوار یا هر وسیله دیگر
- ۴- پخش از رادیو و تلویزیون و وسائل دیگر
- ۵- ترجمه و نشر و تکثیر و عرضه اثر از راه چاپ و نقاشی و عکاسی و گراور و کلیشه و قالب‌ریزی و مانند آن
- ۶- استفاده از اثر در کارهای علمی و ادبی و صنعتی و هنری و تبلیغاتی
- ۷- به کار بردن اثر در فراهم کردن یا پدید آوردن اثرهای دیگری که در ماده دوم این قانون درج شده است

ماده ۸ - کتابخانه‌های عمومی و مؤسسات جمع‌آوری نشریات و مؤسسات علمی و آموزشی که به صورت غیر انتفاعی اداره می‌شوند می‌توانند طبق آیین‌نامه‌ای که به تصویب هیأت وزیران خواهد رسید از اثرهای مورد حمایت این قانون از راه عکسبرداری یا طرق مشابه آن به میزان مورد نیاز و متناسب با فعالیت خود نسخه‌برداری کنند.

خلاصه نظرات متخصصان موضوعی مذکور درباره این مسئله که آیا ایرانداک می‌تواند پیکره زبانی با استفاده از پایان‌نامه‌ها و رساله‌های ثبت شده در ایرانداک بسازد؟ آیا برای این کار نیاز به کسب اجازه از مولفان یا دانشگاه‌ها و مؤسسات طرف قرارداد با ایرانداک دارد؟ به این شرح است:

چنانچه ایرانداک از ساخت پیکره، قصد درآمدزایی نداشته باشد و پیکره را به عنوان یک خدمت عمومی ارائه دهد تا همه امکان دسترسی و استفاده از پیکره را داشته باشند، ممانعت قانونی برای این کار وجود ندارد به شرط آنکه ابتدا، پیش از ساخت پیکره، از دانشگاه‌ها و موسساتی که طرف قرارداد ایرانداک هستند و پایان‌نامه‌ها و رساله‌های دانشجویان آن‌ها در ایرانداک ثبت می‌شود، اجازه‌نامه ساخت پیکره از آن داده‌ها را دریافت کند. این کار مستلزم این است که آن دانشگاه‌ها و مؤسسات قبلاً از مولفان اثر (دانشجویانی که پایان‌نامه و رساله در آن دانشگاه نوشته‌اند) طی تعهدنامه‌ای، اجازه گرفته باشند. مسئله مهم این است که در نوشتن متن قرارداد استفاده از پایان‌نامه‌ها و رساله‌ها در ساخت پیکره، با دانشگاه‌ها و مؤسسات مربوطه، باید دقت نظر لازم صورت پذیرد و مشخصاً در متن قرارداد ذکر شود که دانشگاه یا مؤسسات حق قانونی نمایندگی اثر مولف را دارند و بنابراین، مسوولیت رعایت حقوق مادی و معنوی اثر بر

عهده آن دانشگاه یا موسسه است. به عنوان نمونه، در دانشگاه تهران تعهدنامه‌ای به این مضمون به امضای دانشجو می‌رسد: «کلیه حقوق مادی اثر به دانشگاه تهران تعلق دارد....». به این ترتیب، ایرانداک در حال و آینده، مبرا از هر گونه اتهامی در زمینه نقض حقوق مولف است و کسی حق اعتراض به ایرانداک را ندارد.

اگر ایرانداک نتواند طی شرایط ذکر شده در پاراگراف قبل، از دانشگاه‌ها و موسسات طرف قرارداد اجازه‌نامه ساخت پیکره دریافت کند، باید از خود مولفان (دانشجویان) اجازه گرفت که این کار می‌تواند در قالب قرار دادن بندی در فرم ثبت پایان‌نامه و رساله در ایرانداک ارائه شود که در آن از مولف پرسیده شود: آیا اجازه می‌دهد از پایان‌نامه یا رساله او در ساخت پیکره استفاده شود؟ البته این روش بسیار چالش‌برانگیز خواهد بود، چون طرف قرارداد شخص مولف می‌شود و ممکن است هر فرد شرایط خاصی (مادی و معنوی) را از ایرانداک طلب کند.

راه دیگر می‌تواند این باشد که از وزیر علوم/ وزیر عتف اجازه‌نامه دریافت شود. این راه به شرطی قابل اجرا است که ایشان نیز حق قانونی نمایندگی اثر مولف را داشته باشند یا قدرت اجرایی صادر کردن اجازه‌نامه را داشته باشند.

نکته مهم: در مستندات مربوط به پیکره، باید حتماً ذکر شود که از چه منابعی برای ساخت پیکره استفاده شده است؛ به عنوان نمونه «پایان‌نامه‌ها و رساله‌های دانشگاه‌ها و موسسات.....از تاریخ.....تا.....». همچنین پیشنهاد می‌شود برای اعلام حسن‌نیت، در قسمتی از سایت ثبت پایان‌نامه‌ها و رساله‌ها ذکر شود که ایرانداک از پایان‌نامه‌ها و رساله‌های ثبت شده با دریافت اجازه‌نامه از دانشگاه‌ها و موسسات طرف قرارداد با ایرانداک، در ساخت پیکره استفاده کرده است و پیکره به صورت خدمت عمومی در اختیار همگان است. چون دریافت اجازه‌نامه از تک‌تک مولفان، امکان‌پذیر نبوده است، هر مولفی که پایان‌نامه یا رساله او در ایرانداک ثبت شده است، می‌تواند نام خود را برای ثبت در مستندات منابع پیکره، اعلام کند. که در اینصورت باید قسمتی هم در پیکره برای وارد کردن نام مولفان در نظر گرفته شود.

➤ ابعادی مانند «تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه) و چگونگی تامین منابع مالی، از طریق مشاوره با رئیس پژوهشگاه، مشخص شده است.

اکنون با مشخص بودن ابعاد مالی و حقوقی پروژه ساخت پیکره، پژوهشگاه می‌تواند درباره ساخت پیکره از داده‌های علمی ایرانداک و شروع فاز اول تصمیم‌گیری کند.

۵-۵- خلاصه فصل

در این فصل، پس از معرفی کوتاهی از پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) و شرح کلی فعالیت‌های این پژوهشگاه، به توصیف طراحی مدل امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک و چگونگی اعتباریابی آن پرداخته شد. سپس، مدل نهایی امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک، ارائه شد (جدول ۴-۵) و در نهایت، امکان‌سنجی این مدل در ایرانداک انجام شد؛ به این صورت که هزینه ابعاد و مولفه‌های مدل و چگونگی تصمیم‌گیری درباره ابعاد حقوقی، تامین منابع و فازبندی نیز مشخص گردید (جدول ۵-۵) تا از این رهگذر بهتر بتوان درباره پروژه ساخت پیکره زبانی از داده‌های علمی ایرانداک تصمیم‌گیری کرد.

فصل ششم

نتیجه‌گیری و پیشنهاد برنامه ساخت پیکره زبانی از داده‌های علمی ایرانداک

۶-۱- بحث و نتیجه‌گیری

در پژوهش حاضر پس از بررسی پژوهش‌های پیشین مربوط به ملاحظات و مراحل ساخت پیکره، ابتدا شاخص‌های اصلی ساخت پیکره زبانی (فارغ از اینکه پیکره از چه نوع باشد) استخراج شد. این شاخص‌ها شامل: نمونه‌گیری، نمایندگی، توازن، اندازه، نوع پیکره و یک‌دستی است که مولفه‌های مربوط به هر کدام در جدول زیر قابل مشاهده است.

شاخص	شرح
یک‌دستی	کنار نهادن متون بدون پیوند با هدف پیکره
گونه پیکره	عمومی / تخصصی
اندازه	شمار واژه‌های پیکره
توازن	۱. چه زبانی و چه اندازه به کار رود ۲. متون چه انداز تخصصی باشند ۳. متون چه اندازه رسمی یا غیررسمی باشند ۴. گزینش یک پیکره مبنا
نمایندگی	۱. اندازه ژانرها و گونه‌های متون در پیکره ۲. چگونگی گزینش متون در پیوند با هر ژانر
نمونه‌گیری	۱. گرایش / سوگیری سازندگان پیکره ۲. معیارهای گزینش نمونه‌ها: □ متن (گفتاری / نوشتاری / ...) □ واحد نمونه‌گیری (کتاب / مجله / مقاله / ...) □ زمینه متن (دانشگاهی / عمومی / ...) □ زبان متن (معیار / گونه ویژه زبانی) □ مکان (گونه زبانی وابسته به مکانی ویژه) □ زمان (تاریخ انتشار متن) ۳. روش نمونه‌گیری (تصادفی ساده / لایه لایه) ۴. اندازه نمونه (متن کامل / بریده متن) ۵. ژانر متون (ادبیات / کتاب درسی / قوانین و مقررات / ...)

پس از این مرحله، مراحل ساخت پیکره متنی و پیکره گفتاری، که رایج‌ترین نوع پیکره هستند، توضیح داده شد. فرآیند ساخت پیکره متنی شامل: طراحی ساختار و شکل کلی پیکره، گزینش متون، پیش-پردازش متون و حاشیه‌نویسی است. مراحل ساخت پیکره گفتاری هم شامل: طراحی ساختار و شکل کلی پیکره، جمع‌آوری داده‌ها، آوانویسی، نمایش و حاشیه‌نویسی و دسترسی است.

از آنجایی که هدف نهایی در پژوهش‌های بعدی، استفاده از داده‌های مکتوب علمی ایرانداک برای ساخت پیکره است، باید امکان‌سنجی ساخت پیکره متنی انجام شود تا ابزار، تیم‌های تخصصی و هزینه‌های لازم مشخص شود. در این راستا، برای امکان‌سنجی ساخت پیکره متنی، ابتدا چارچوب مفهومی این امکان‌سنجی با در نظر گرفتن ابعاد و معیارهای مربوطه ساخته شد؛ در این راستا از روش دلفی استفاده شد. به این صورت که ابتدا پرسش‌نامه‌ای تهیه شد و در اختیار سه متخصص موضوعی قرار گرفت تا نظرات آن‌ها درباره ابعاد و مولفه‌های امکان‌سنجی ساخت پیکره متنی که از بررسی ادبیات مربوطه به دست آمده-اند، دریافت شود. پس از اعمال نظرات آن‌ها، در دور اول دلفی پرسش‌نامه نهایی به ده متخصص موضوعی تحویل داده شد تا پاسخ دهند. این پرسش‌نامه به متخصصین اجازه می‌داد، غیر از مولفه‌های ذکر شده در پرسش‌نامه، اگر مولفه‌های دیگری نیز وجود دارد، ذکر کنند. نظرات آن‌ها طی پرسش‌نامه دیگری در دور دوم دلفی در اختیار همان ده متخصص موضوعی قرار گرفت تا پاسخ دهند. با در نظر گرفتن نظرات متخصصان درباره مولفه‌های امکان‌سنجی ساخت پیکره، مدل امکان‌سنجی ساخت پیکره متنی و ابعاد و مولفه‌های مربوطه طراحی شد که در جدول زیر قابل مشاهده است:

مدل امکان‌سنجی ساخت پیکره متنی

ابعاد	مولفه‌ها
۱. فنی	۱-۱۲. بررسی نیاز به زبان‌شناس همگانی / عمومی در تیم ساخت پیکره
	۱-۱۳. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره
	۱-۱۴. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره
	۱-۱۵. بررسی نیاز به پشتیبانی برون‌سازمانی از پیکره
	۱-۱۶. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره
	۱-۱۷. انتخاب نوع متون، به عنوان گامی از ساخت پیکره
	۱-۱۸. تعیین حجم متون، به عنوان گامی از ساخت پیکره
	۱-۱۹. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره
	۱-۲۰. حاشیه‌نویسی (برچسب‌گذاری و...)، به عنوان گامی از ساخت پیکره
	۱-۲۱. بررسی نیاز به رابط کاربر (User interface) برای کار کردن کاربران با پیکره
	۱-۲۲. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند
۲. اقتصادی	۲-۱۳. برآورد هزینه متخصصان ساخت پیکره

۱۴-۲. برآورد هزینه برنامه‌نویسی رایانه	
۱۵-۲. برآورد هزینه حاشیه‌نویسی دستی	
۱۶-۲. برآورد هزینه آموزش کاربران	
۱۷-۲. برآورد هزینه خرید نرم‌افزار	
۱۸-۲. برآورد هزینه وسایل و تجهیزات اداری (رایانه، میز، صندلی و....)	
۱۹-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و...)	
۲۰-۲. برآورد هزینه نگهداری زیرساخت (سرور، شبکه، و...)	
۲۱-۲. برآورد هزینه ساخت ابزار گردآوری متن	
۲۲-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره	
۲۳-۲. تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)	
۲۴-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پایشگر / بروزشونده	
۱-۳. برآورد زمان‌بندی ساخت پیکره	۳. زمان‌بندی
۲-۳. فازبندی زمان ساخت پیکره	
۴-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره	۴. قانونی / حقوقی
۵-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره	
۶-۴. بررسی حقوق مادی پیکره (پس از ساخت)	
۳-۵. بررسی پذیرش و آمادگی سازمان در کاربرد پیکره	۵. عملیاتی
۴-۵. بررسی دستاوردهای سازمان با ساخت پیکره	
۱-۶. پیش‌بینی منابع تأمین مالی برای ساخت پیکره	۶. تأمین مالی
۳-۷. شناخت کاربران پیکره	۷. بازاریابی
۴-۷. آزمون عملکرد پیکره پیش از عرضه به بازار	

پس از طراحی مدل کلی امکان‌سنجی ساخت پیکره متنی، در فصل پنجم به توصیف چگونگی طراحی مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک پرداخته شد؛ برای طراحی مدل مذکور پنلی متشکل از رئیس پژوهشگاه و نه متخصص علمی و اجرایی شاغل در ایرانداک تشکیل شد. پیش از برگزاری پنل، ابتدا پرسش‌نامه‌ای تهیه شد که حاوی همان ابعاد و مولفه‌هایی است که در مدل کلی امکان‌سنجی ساخت پیکره متنی (فصل چهارم، جدول ۴-۱۵ و شکل ۴-۶) آورده شده است، این پرسش‌نامه در پنل در اختیار اعضای پنل قرار گرفت. پاسخ‌های شرکت‌کنندگان در پنل مورد تجزیه و تحلیل آماری قرار گرفت و در نهایت مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک طراحی شد که در جدول زیر قابل مشاهده است.

مدل امکان‌سنجی ساخت پیکره متنی از داده‌های علمی ایرانداک

ابعاد	مؤلفه‌ها
۱. فنی	۱۱-۱. بررسی نیاز به زبان‌شناس همگانی/ عمومی در تیم ساخت پیکره ۱۲-۱. بررسی نیاز به زبان‌شناس رایانشی در تیم ساخت پیکره ۱۳-۱. بررسی نیاز به برنامه‌نویس رایانه‌ای در تیم ساخت پیکره ۱۴-۱. طراحی زیربنای پیکره (چارچوب و ساختار)، به عنوان گامی از ساخت پیکره ۱۵-۱. انتخاب نوع متون، به عنوان گامی از ساخت پیکره ۱۶-۱. تعیین حجم متون، به عنوان گامی از ساخت پیکره ۱۷-۱. پیش‌پردازش متون، به عنوان گامی از ساخت پیکره ۱۸-۱. حاشیه‌نویسی (برجسب‌گذاری و...)، به عنوان گامی از ساخت پیکره ۱۹-۱. بررسی نیاز به رابط کاربر (User interface) برای کار کردن کاربران با پیکره ۲۰-۱. بررسی نیاز به همکاری پژوهشگرانی که تجربه ساخت پیکره دارند
۲. اقتصادی	۸-۲. برآورد هزینه متخصصان ساخت پیکره ۹-۲. برآورد هزینه برنامه‌نویسی رایانه ۱۰-۲. برآورد هزینه حاشیه‌نویسی دستی ۱۱-۲. برآورد هزینه خرید نرم‌افزار ۱۲-۲. برآورد هزینه اولیه تأمین زیرساخت (سرور، شبکه، و...) ۱۳-۲. برآورد هزینه ساخت ابزار کار با داده‌های پیکره ۱۴-۲. برآورد هزینه بروز کردن داده‌های پیکره در پیکره‌های پیش‌گزر/ بروزشونده
۳. زمان‌بندی	۱-۳. برآورد زمان‌بندی ساخت پیکره ۲-۳. فازبندی زمان ساخت پیکره
۴. قانونی / حقوقی	۳-۴. بررسی الزام‌های قانونی و حقوقی در ساخت پیکره ۴-۴. بررسی اجازه‌نامه کاربرد متون در ساخت پیکره
۵. تأمین مالی	۱-۵. پیش‌بینی منابع تأمین مالی برای ساخت پیکره
۶. بازاریابی	۳-۶. شناخت کاربران پیکره ۴-۶. آزمون عملکرد پیکره پیش از عرضه به بازار

پس از این مرحله، امکان‌سنجی ساخت پیکره از داده‌های علمی ایرانداک صورت پذیرفت که در آن هزینه‌های ثابت و متغیر مربوط به هر مؤلفه و تصمیم‌گیری درباره مسائل حقوقی، زمان‌بندی و..... مشخص شد (جدول ۵-۵). انجام این امکان‌سنجی به ایرانداک کمک خواهد کرد تا بهتر بتواند درباره پروژه ساخت پیکره زبانی از داده‌های علمی ایرانداک، تصمیم‌گیری و برنامه‌ریزی نماید.

۶-۲- مقایسه یافته‌ها با پژوهش‌های پیشین

همانگونه که در بخش پیشینه پژوهش ذکر شد، تاکنون پژوهشی درباره امکان‌سنجی ساخت پیکره زبانی انجام نشده است. بنابراین، ارائه مدلی فراگیر برای ساخت پیکره، مهم‌ترین مزیت پژوهش حاضر نسبت به پژوهش‌های مشابه در حوزه امکان‌سنجی است. در پژوهش حاضر، هفت بعد از امکان‌سنجی (شامل: بعد فنی، بعد اقتصادی، بعد قانونی/حقوقی، بعد زمان‌بندی، بعد عملیاتی، بعد بازاریابی و بعد تامین منابع) مورد بررسی قرار گرفته است که از این نظر تا حدی شبیه ابعاد مورد بررسی در پژوهش‌های موخری و رئی (۲۰۱۷) و بوس و همکاران (۲۰۱۴) است که پنج بعد اصلی امکان‌سنجی پروژه را مورد بحث و بررسی قرار داده‌اند: امکان‌سنجی فنی، امکان‌سنجی اقتصادی، امکان‌سنجی قانونی/حقوقی، امکان-سنجی عملیاتی و امکان‌سنجی زمان‌بندی. همچنین با در نظر گرفتن عناصر یک امکان‌سنجی خوب و موفقیت‌آمیز (اطمینان از صحت اطلاعات)، پژوهش حاضر در جایگاه خوبی قرار دارد، زیرا بر اساس مرور ادبیات مربوطه و استفاده از نظر متخصصان موضوعی این حوزه تهیه شده است. همچنین پژوهش حاضر مانند پژوهش نجفی‌خواه و همکاران (۱۳۹۵)، کاربردی و با استفاده از پرسش‌نامه، انجام شده است. اگر پژوهشی داخلی یا خارجی در زمینه امکان‌سنجی ساخت پیکره انجام شده بود، می‌توانستیم نتایج پژوهش حاضر را با آن‌ها مقایسه کنیم، پژوهش‌های داخلی یا خارجی در حوزه ساخت پیکره، به شاخص‌های کلی ساخت پیکره و چالش‌های مربوطه که بخشی از آن، چالش‌های خط و زبان است، پرداخته شده است و این اولین بار است که پژوهشی در زمینه امکان‌سنجی ساخت پیکره انجام می‌شود.

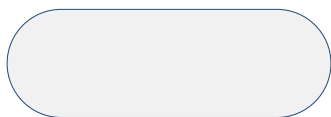
۶-۳- پیشنهاد برنامه ساخت پیکره زبانی از داده‌های علمی ایرانداک

امکان‌سنجی‌ها راه را برای کسانی که قصد ساخت انواع پیکره دارند، هموار خواهد ساخت. به طور کلی، پیشنهاد می‌شود پژوهشگران و سازمان‌های تأمین‌کننده مالی، پیش از آغاز هر پروژه ساخت پیکره متنی، نخست امکان‌سنجی آن را به انجام دهند تا بتوانند با شناخت و آگاهی بیشتری آن را به سرانجام برسانند. این شناخت و آگاهی، به ویژه در پروژه‌های بزرگ می‌تواند اثربخشی و کارایی را در پروژه به ارمغان آورد.

دست‌یابی به هدف پروژه مستلزم در دست داشتن چارچوب مفهومی و انجام فعالیت‌هایی است که در نهایت منتهی به ساخت پیکره می‌شود.

امكانسنجى ساخت پيكره زباني از داده‌هاى علمى پژوهشگاه علوم و فناورى اطلاعات ايران (ايرانداك) □ ۱۴۳

۱-۳-۶- چارچوب مفهومی ساخت پیکره متنی از داده‌های علمی ایرانداک



۶-۳-۲- فعالیت‌ها

۶-۳-۲-۱- دریافت اجازه‌نامه

پیش از شروع پروژه ساخت پیکره از داده‌های علمی ایرانداک، باید ابتدا از دانشگاه‌ها و موسسات طرف قرارداد ایرانداک (منظور دانشگاه‌ها و موسساتی هستند که پایان‌نامه‌ها و رساله‌های دانشجویانشان در ایرانداک ثبت می‌گردد)، اجازه‌نامه ساخت پیکره از داده‌های علمی آن‌ها دریافت گردد. اگر به هر دلیلی این امکان فراهم نباشد، باید از خود مولفان (دانشجویان) هنگام ثبت پایان‌نامه یا رساله در ایرانداک، اجازه‌نامه استفاده از پایان‌نامه و رساله برای ساخت پیکره دریافت شود، که در اینصورت باید صبر کرد تا تعداد اجازه‌نامه‌ها به حد معین و معقولی برسد تا بتوان از آن پایان‌نامه‌ها و رساله‌ها دارای مجوز، در ساخت پیکره استفاده کرد. ضمناً در مستندات پیکره حتماً فهرستی از منابع استفاده شده در ساخت پیکره وارد شود.

۶-۳-۲-۲- تهیه تجهیزات و مکان انجام پروژه

پس از اطمینان از عدم وجود منع قانونی برای ساخت پیکره، باید مکان انجام پروژه (یک اتاق با تجهیزات لازم (میز، صندلی، رایانه و....) مشخص شود تا هم برنامه‌نویس در آنجا مستقر شود و هم در صورت لزوم جلسات در آنجا برگزار گردد.

وضعیت ایرانداک: در حال حاضر ایرانداک توانایی تهیه یک اتاق و تجهیزات لازم برای انجام پروژه ساخت پیکره را دارد.

۶-۳-۲-۳- تشکیل تیم ساخت پیکره

تیم ساخت پیکره متشکل از زبان‌شناس عمومی، زبان‌شناس رایانشی، برنامه‌نویس رایانه‌ای و پژوهشگر/پژوهشگرانی است که تجربه ساخت پیکره دارند. در این پروژه بهتر است از یک برنامه‌نویس رایانه‌ای حرفه‌ای که تجربه ساخت پیکره هم دارد برای حضور در تیم اجرایی دعوت شود که خود توانایی طراحی و ساخت نرم‌افزارهای مورد نیاز را داشته باشد. پیشنهاد می‌شود هزینه‌ای که برای برنامه‌نویسی پروژه ساخت پیکره در نظر گرفته شده است یا به صورت ماهانه (پس از ارائه گزارش پیشرفت پروژه) یا طی سه مرحله (مرحله اول: هنگام شروع پروژه، مرحله دوم: پس از طراحی شکل کلی پیکره و پیش‌پردازش متون و مرحله سوم: پس از اتمام پروژه) به ایشان پرداخت گردد. کارهایی که برنامه‌نویس رایانه‌ای انجام خواهد

داد شامل: طراحی زیربنای پیکره، پیش‌پردازش متون، طراحی رابط کاربر و ساخت ابزار کار با داده‌های پیکره، وارد کردن متون و پشتیبانی نرم‌افزاری/برنامه‌نویسی است (هزینه همه کارها به غیر از پشتیبانی، ثابت است). جزئیات هزینه‌های مربوط به متخصصان دیگر در جدول ۵-۵، آورده شده است. همچنین در تیم ساخت پیکره از زبان‌شناسان رایانشی دعوت به همکاری شود که تجربه ساخت پیکره هم دارند. در این صورت، دیگر نیازی نیست هزینه‌ای جدا برای همکاری پژوهشگرانی که تجربه ساخت پیکره دارند، در نظر گرفته شود. بنابراین، گویی دو کار با هم انجام می‌شود و در هزینه‌ها کمی صرفه‌جویی می‌شود.

وضعیت ایرانداک: به دلیل حضور زبان‌شناس عمومی در میان اعضای هیات علمی ایرانداک، نیازی به دعوت زبان‌شناس عمومی از بیرون از پژوهشگاه نیست. اما از آنجائیکه داخل پژوهشگاه کسی تجربه ساخت پیکره ندارد، باید از پژوهشگرانی خارج از پژوهشگاه که تجربه ساخت پیکره دارند دعوت به همکاری کرد. برنامه‌نویس رایانه‌ای حرفه‌ای که تجربه کار با پیکره دارد نیز از بیرون از پژوهشگاه دعوت می‌شود.

۶-۳-۲-۴- فازبندی پروژه ساخت پیکره

پروژه ساخت پیکره در دو فاز انجام شود. پیشنهاد می‌شود ابتدا در پژوهشی، فاز اول اجرا شود تا ابعاد کار به درستی معلوم گردد و سپس، با توجه به حجم متون و بودجه پژوهشگاه، فاز دوم عملیاتی گردد. شرح دو فاز به صورت زیر است:

فاز اول: تشکیل تیم ساخت پیکره (متشکل از زبان‌شناس عمومی، زبان‌شناس رایانشی، برنامه‌نویس رایانه‌ای و پژوهشگر/پژوهشگرانی است که تجربه ساخت پیکره دارند)، طراحی زیربنای پیکره و پیش‌پردازش متون. در پایان فاز اول، در واقع پیکره ساخته شده است، اما پیکره‌ای است حاشیه‌نویسی نشده و برچسب‌گذاری داده‌ها انجام نشده است. تا همین مرحله هم می‌توان پایان پروژه ساخت پیکره را اعلام کرد، زیرا از پیکره حاشیه‌نویسی نشده هم می‌توان استفاده کرد. به عنوان نمونه، از این نوع پیکره‌ها می‌توان برای انواع طبقه‌بندی اطلاعات استفاده کرد. البته می‌توان در همین مرحله برخی از انواع حاشیه‌نویسی‌های خودکار را هم وارد کرد. اما برای بالا بردن کارایی پیکره، در فاز اول تصمیم می‌گیریم کارایی پیکره را با انواع حاشیه‌نویسی بالا ببریم که در آن صورت باید حجم متون و نوع حاشیه‌نویسی معلوم شده باشد تا بتوان آن نوع حاشیه‌نویسی را ساخت یا دستی وارد و کنترل کرد.

فاز دوم: تشکیل تیم حاشیه‌نویسی و انجام حاشیه‌نویسی است.

۶-۳-۲-۵- هماهنگی با بخش فنی ایرانداک

همزمان با شروع ساخت پیکره، با بخش فنی ایرانداک هماهنگی‌های لازم انجام شود و طی جلسه/جلساتی مشخص شود چه زمانی و چگونه زیرساخت‌های ساخت پیکره تامین و سپس، پشتیبانی شود. پیشنهاد می‌شود در این زمینه، بخش فنی فردی را برای همکاری معرفی کند.

وضعیت ایرانداک: در حال حاضر ایرانداک توانایی تامین و سپس، پشتیبانی زیرساخت‌های ساخت پیکره را دارد.

۶-۳-۳-۳- ابعاد، زمان‌بندی و هزینه‌های ساخت پیکره متنی از داده‌های علمی ایرانداک

ابعاد امکان‌سنجی	تصمیم‌گیری یا مشخص نمودن هزینه مربوطه (به ریال)	نوع هزینه (ثابت یا متغیر)	جمع هزینه‌های ثابت به ریال	جمع هزینه‌های متغیر به ریال (سالانه)
۱- بعد فنی			700,000,000 (هفتاد میلیون تومان)	560,000,000 (پنجاه و شش میلیون تومان)
۱-۱- مشاوره با زبان‌شناس همگانی/عمومی	50,000,000 (پنج میلیون تومان)	ثابت	شامل:	هزینه متغیر شامل موارد زیر است:
۱-۲- مشاوره با زبان‌شناس رایانشی	100,000,000 (ده میلیون تومان)	ثابت	۱. مشاوره با زبان‌شناس همگانی/عمومی (در فاز اول و دوم)	۱. هزینه تامین و نگهداری زیرساخت (سرور، شبکه و....)
۱-۳- همکاری برنامه‌نویس رایانه‌ای	500,000,000 (پنجاه میلیون تومان)	برای فاز اول: ثابت (وظایف ایشان شامل: طراحی زیربنای پیکره (پانزده میلیون تومان)، پیش‌پردازش متون (ده میلیون تومان)، طراحی رابط کاربر و ساخت	۲. مشاوره با زبان‌شناس رایانشی که تجربه ساخت پیکره هم داشته باشند (در فاز اول و دوم)	۲. هزینه بروز کردن داده‌های

پیکره در پیکره- های پایشگر/پروزشوند ه ۳. پشتیبانی نرم- افزایی / برنامه- نویسی	اول). وظایف: الف. طراحی زیربنای پیکره ب. پیش‌پردازش متون پ. طراحی رابط کاربر و ساخت ابزار کار با داده‌های پیکره (مبلغ پنج میلیون تومان هم برای مسائل پیش- بینی نشده که ممکن است هنگام کار به وظایف برنامه‌نویس اضافه شود، در نظر گرفته شده است) ۴. خرید نرم‌افزار	ابزار کار با داده‌های پیکره (بیست میلیون است) و پشتیبانی نرم- افزایی (برنامه‌نویسی) که البته هزینه آن متغیر است و سالانه محاسبه می- گردد. برای فاز دوم، باید پس از تعیین حجم متون، برآورد هزینه صورت پذیرد.		
		ثابت	اگر از متخصصان زبان‌شناسی رایانشی که تجربه ساخت پیکره دارند، استفاده شود، هزینه آن‌ها در بخش «مشاوره با زبان‌شناس رایانشی» دیده شده است و دیگر نیازی نیست در این قسمت وارد شود. اما اگر زبان‌شناس/زبان‌شناسان رایانشی حاضر در تیم، تجربه ساخت پیکره نداشته باشند، می‌توان هزینه‌ای حدود 100,000,000 (ده میلیون تومان) را برای مشاوره با پژوهشگرانی که تجربه ساخت پیکره دارند، در نظر گرفت.	۴-۱- همکاری پژوهشگرانی که تجربه ساخت پیکره دارند
		ثابت	150,000,000 (پانزده میلیون تومان) (این هزینه در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	۵-۱- طراحی زیربنای پیکره (چارچوب و ساختار)
			با مشاوره متخصص مشخص خواهد شد	۶-۱- انتخاب نوع متون
			با مشاوره متخصص مشخص خواهد شد	۷-۱- تعیین حجم متون
		ثابت	100,000,000	۸-۱- پیش‌پردازش متون

			(ده میلیون تومان) (این هزینه در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	
			به حجم داده‌های پیکره بستگی دارد و باید در فاز دوم که حجم پیکره مشخص شده است، این هزینه محاسبه شود.	۹-۱- حاشیه‌نویسی (برچسب‌گذاری و....)
		ثابت	150,000,000 (پانزده میلیون تومان) (این بخش در هزینه برنامه‌نویس رایانه‌ای دیده شده است)	۱۰-۱- رابط کاربر (user interface)
				۲- بعد اقتصادی
			در بعد فنی دیده شده است	۱-۲- هزینه همکاری متخصصان
			در بعد فنی دیده شده است	۲-۲- هزینه برنامه‌نویس رایانه‌ای
			به حجم داده‌های پیکره بستگی دارد	۳-۲- هزینه حاشیه‌نویسی دستی
		ثابت	50,000,000 (پنج میلیون تومان)	۴-۲- هزینه خرید نرم‌افزار
		متغیر (سالانه)	360,000,000 (ماهانه سه میلیون تومان که سالانه سی و شش میلیون تومان، می‌شود؛ تامین و نگهداری زیرساخت به بخش فنی پژوهشگاه مربوط می‌شود)	۵-۲- هزینه تامین و نگهداری زیرساخت (سرور، شبکه و....)
		ثابت	50,000,000 (پنج میلیون تومان) (این بخش در واقع بخشی از کار طراحی رابط کاربر است، اما چون در مدل امکان‌سنجی به عنوان بخشی جدا در نظر گرفته شده، هزینه هم جدا آورده شده است)	۶-۲- هزینه ساخت ابزار کار با داده‌های پیکره

		متغیر (سالانه)	100,000,000 (ده میلیون تومان)	۷-۲- هزینه بروز کردن داده‌های پیکره در پیکره-های پایشگر/بروزشونده
			ایرانداک تصمیم دارد این پیکره را به رایگان در اختیار کاربران قرار دهد.	۸-۲- تصمیم‌گیری درباره چگونگی دسترسی کاربران به پیکره (رایگان یا با پرداخت هزینه)
		متغیر (سالانه)	بخش پشتیبانی شامل دو قسمت است: ۱. پشتیبانی از مسائل کاربران: در این بخش باید کسی در پژوهشگاه همکاری کند تا اگر نیازی به پاسخ‌گویی به درخواست یا ایمیل کاربران باشد، این کار را انجام دهد. بنابراین، پژوهشگاه می‌تواند پس از ساخت پیکره، درباره این بخش، تصمیم‌گیری کند. ۲. پشتیبانی نرم‌افزاری / برنامه‌نویسی: اگر نیازهای جدیدی از طرف کاربران مطرح گردد و نیاز به ایجاد تغییر یا اضافه کردن feature هایی به پیکره باشد، برنامه‌نویس رایانه‌ای تیم ساخت پیکره می‌تواند این کارها را انجام دهد که هزینه سالانه‌ای می‌توان برای آن در نظر گرفت. 100,000,000 (ده میلیون تومان)	۹-۲- پشتیبانی
				۳- بعد زمان‌بندی
			می‌توان پیکره را در دو فاز دید که طی یک الی دو سال پیکره کلی ساخته شود.	۳-۱- برآورد زمان‌بندی ساخت پیکره
	فاز اول: هزینه ثابت (هفتاد میلیون تومان)	فاز اول:	تشکیل تیم ساخت پیکره، طراحی زیربنای پیکره، پیش‌پردازش متون و	۳-۲- فازبندی ساخت پیکره زمان

		وارد کردن متون در پیکره (۹ الی ۱۲ ماه)	فاز دوم: تشکیل تیم حاشیه‌نویسی، انجام انواع حاشیه‌نویسی‌هایی که در فاز اول درباره آن‌ها تصمیم‌گیری شده است	
		فاز دوم: مشخص نیست و به نوع حاشیه‌نویسی و حجم متون بستگی دارد.		
				۴- بعد قانونی/حقوقی
			جنبه قانونی ساخت پیکره از داده‌های ایرانداک از طریق مشاوره با متخصصان حقوقی مشخص شده است.	۴-۱- بررسی الزام‌های قانونی و حقوقی در ساخت پیکره
			ایرانداک اجازه ساخت پیکره از داده‌های علمی موجود در ایرانداک (پایان-نامه‌ها، رساله‌ها و...) را دارد. اما پیش از ساخت پیکره باید اجازه‌نامه انجام این کار را داشته باشد. اجازه‌نامه را می‌تواند از دانشگاه‌ها و موسسات طرف قرارداد ایرانداک دریافت کند. یا از خود مولفان اثر دریافت کند.	۴-۲- بررسی اجازه‌نامه کاربرد متون در ساخت پیکره
				۵- بعد تأمین مالی
			اکنون تأمین مالی پیکره از منابع پژوهشگاه است و منبع دیگری برای آن پیشنهاد نشده است.	۵-۱- پیش‌بینی منابع تأمین مالی برای ساخت پیکره
				۶- بعد بازاریابی
			کاربران پیکره، نوعاً پژوهشگرانی هستند که در حوزه زبان‌شناسی پیکره‌ای، عمومی، رایانشی و هوش مصنوعی فعالیت دارند. البته بسته به موضوع پژوهش، ممکن است پژوهش‌گران دیگر نیز از پیکره استفاده کنند.	۶-۱- شناخت کاربران پیکره
			بدیهی است، پیکره پیش از اینکه در دسترس کاربران قرار گیرد، عملکرد آن مورد ارزیابی قرار گیرد.	۶-۲- آزمون عملکرد پیکره پیش از عرضه به بازار

--	--	--	--	--

پیشنهادهای کلی:

- در ساخت پیکره‌های بزرگ برای زبان فارسی (مانند پیکره علمی از داده‌های ایرانداک)، پیشنهاد می‌شود که بر پایه ابعاد و مؤلفه‌های امکان‌سنجی، نخست پیکره‌ای کوچک ساخته شود؛ سپس اندازه داده‌ها افزایش یابد.
- می‌توان پروژه ساخت پیکره را ابتدا از مجله منتسب به ایرانداک (پژوهش‌نامه پردازش و مدیریت اطلاعات) شروع کرد.
- پژوهش‌های کاربردی برای ساخت یا بهبود عملکرد ابزارهایی مانند خطایاب، تشخیص‌دهنده گروه‌های نحوی، پارسر، ابزار نرمال‌سازی متن و..... و انواع حاشیه‌نویسی گفتمانی، کاربردی، سبکی، نحوی و..... برای زبان فارسی از دیگر کارهایی است که نیاز به انجام آن‌ها هست. این پژوهش‌ها می‌توانند هم‌زمان با آغاز پروژه ساخت پیکره انجام پذیرند.
- پیشنهاد می‌شود این مدل با روش اقدام‌پژوهی^۱ در یک پروژه ساخت پیکره متنی به کار رود تا برای آن، اعتبار تجربی نیز فراهم شود.
- بررسی دقیق‌تر برخی از مؤلفه‌های مدل نیز می‌تواند جداگانه موضوع پژوهش باشند. برای نمونه «الزام‌های قانونی و حقوقی در ساخت پیکره» در بُعد قانونی حقوقی یا «پذیرش و آمادگی سازمان در کاربرد پیکره» در بُعد عملیاتی، می‌توانند در پژوهش‌هایی جداگانه بررسی و راهکارها و مدل‌های آن‌ها بررسی و تبیین شوند.

¹ action research

فهرست منابع فارسی

- ادیب‌منش، مرزبان، محمدعلی عسگری، و نعمت‌الله موسی‌پور. ۱۳۹۰. امکان‌سنجی اجرای برنامه‌ریزی درسی مدرسه محور در دبیرستان‌های شهر کرمانشاه. *مجله پژوهش‌های برنامه درسی* ۱(۲): ۸۱-۱۰۴.
- بی‌جن‌خان، محمود. ۱۳۸۳. نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای. *مجله زبان‌شناسی* ۱۹(۲): ۶۷-۴۸.
- پورفرج، اکبر، مهرداد کرمی، زهرا نکویی، و زهرا طالب‌بیدختی. ۱۳۹۱. امکان‌سنجی فرصت‌های سرمایه‌گذاری صنعت گردشگری. *مطالعه موردی: شهرک تفریحی-توریستی طارم استان زنجان. فصل‌نامه مطالعات مدیریت صنعتی* ۹(۲۵): ۱۸۹-۲۱۴.
- تابع بردبار، علیرضا. ۱۳۹۳. *ایجاد پیکره زبانی موازی به وسیله پیکره‌های قیاس‌پذیر*. کارشناسی ارشد. شیراز: دانشگاه شیراز.
- جواهری، حسن. ۱۳۸۳. *مطالعات امکان‌سنجی پروژه‌های تفریحی و گردشگری. اولین کنفرانس بین‌المللی مدیریت پروژه، تهران، گروه پژوهشی آریانا*.
- حسین‌نژاد، شادی، یاسر شکفته، و طاهره امامی آزادی. ۱۳۹۶. پیکره اعلام: یک پیکره استاندارد واحدهای اسمی برای زبان فارسی. *فصل‌نامه پردازش علائم و داده‌ها* ۳(۳۳): ۱۴۰-۱۲۷.
- حسینی معصوم، سیدمحمد، و مصطفی غضنفری. ۱۳۹۱. کاربرد پیکره‌های زبانی پویا در ترجمه اینترنتی متون فارسی به انگلیسی؛ بررسی عوامل مؤثر در بروز خطا. *پژوهش‌های زبان‌شناسی تطبیقی* ۲(۳): ۱۴۷-۱۲۹.
- دشتبانی، شکوفه. ۱۳۹۱. *ساخت پیکره متنی فارسی حوزه فاوا*. همدان: دانشگاه بوعلی سینا. دانشکده فنی و مهندسی.
- سالاری، سمیه. ۱۳۹۵. *امکان‌سنجی پیاده‌سازی متورینگ در صلاحیت حرفه‌ای معلمان دوره ابتدایی*. تهران: دانشگاه علامه طباطبائی. دانشکده علوم تربیتی و روانشناسی.
- شکری، فاطمه. ۱۳۹۴. *امکان‌سنجی متورینگ تلفیقی و چگونگی پیاده‌سازی آن میان اعضای هیئت علمی دانشگاه علامه طباطبائی*. تهران: دانشگاه علامه طباطبائی. دانشکده روانشناسی و علوم تربیتی.

امکان‌سنجی ساخت پیکره زبانی از داده‌های علمی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) □ ۱۵۳

- صراف‌زاده، اصغر، و وحیده علیپور. ۱۳۹۰. امکان‌سنجی ارائه محصول بیمه‌ای موجود به صورت الکترونیک در شرکت‌های بیمه پارسیان، آسیا، البرز، رازی، توسعه، سینا و معلم. فصل‌نامه مدیریت توسعه و تحول ۷. ۴۸-۳۹.
- عاصی، مصطفی. ۱۳۸۴. گزارش کوتاهی از شکل‌گیری پایگاه داده‌های زبان فارسی در اینترنت. مجله پژوهشگران شماره ۲. صفحه ۱۳.
- <https://hawzah.net/fa/Magazine/View/6280/6282/69701>
- عاصی، مصطفی. ۱۳۸۲. از پیکره زبانی تا زبان‌شناسی پیکره‌ای. مجموعه مقالات پنجمین کنفرانس زبان‌شناسی. دانشگاه علامه طباطبائی.
- علایی ابودر، الهام، و محمود بی‌جن‌خان. ۱۳۹۲. عمق خط فارسی. پژوهش‌های زبانی (مجله سابق دانشکده ادبیات و علوم انسانی دانشگاه تهران) ۴(۱): ۱۹-۱.
- علیدوستی، سیروس. ۱۳۸۵. روش دلفی: مبانی، مراحل و نمونه‌هایی از کاربرد. فصلنامه علمی-ترویجی مدیریت و توسعه ۸(۳۱): ۲۳-۸.
- قاسمی زاده، بهرنگ، سعید رحیمی، مرتضی سلاریان، و علی بهاری‌سلیم. ۱۳۸۵. گزارشی از ساخت نخستین پیکره چندزبانه برای زبان فارسی. دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران. تهران: دانشگاه شهید بهشتی.
- کامیابی گل، عطیه، الهام اخلاقی باقوجری، احسان عسگریان، و هانیه حبیبی. ۱۳۹۷. استخراج اطلاعات از پیکره زبانی: معرفی پیکره مقالات علمی-پژوهشی دانشگاه فردوسی مشهد. مقالات آماده انتشار. فصل‌نامه کتابداری و اطلاع‌رسانی. سازمان کتابخانه‌ها، موزه‌ها و مرکز اسناد آستان قدس رضوی.
- کوهستانی، منوچهر. ۱۳۸۹. بررسی خطاهای املائی و نگارشی در وبلاگ‌های فارسی و ماهیت زبان‌شناختی آن‌ها. تهران: دانشگاه تهران.
- لطفعلی‌پور، محمدرضا، و سکینه اسلامی گیسکی. ۱۳۸۶. ارزیابی هزینه-فایده و تحلیل حساسیت مجتمع مس سرچشمه. فصل‌نامه بررسی‌های اقتصادی ۴(۳): ۹۹-۸۳.
- مدرس خیابانی، شهرام. ۱۳۸۴. بررسی مسئله باهم‌آیی واژگانی در زبان فارسی. مجموعه مقالات نخستین همایش انجمن زبان‌شناسی ایران. به کوشش دکتر مصطفی عاصی، دانشگاه تهران.

- مرادی، مهدی، و بهرام وزیرنژاد. ۱۳۹۱. ساخت پیکره‌های نشانه‌گذاری شده با رویکرد وب به عنوان پیکره. دومین هم‌اندیشی زبان‌شناسی رایانشی. به همت انجمن زبان‌شناسی ایران و پژوهشگاه علوم انسانی و مطالعات فرهنگی. ۱۵۳-۱۵۸.
- مرادی، مهدی، و محمد بحرانی. ۱۳۹۴. تشخیص خودکار جنسیت نویسنده در متون فارسی. پردازش علائم و داده‌ها ۱۲ (۴): ۸۳-۹۴.
- میرزایی، آزاده، و امیرسعید مولودی. ۱۳۹۳. نخستین پیکره نقش‌های معنایی در زبان فارسی. علم زبان ۲ (۳): ۲۹-۴۸.
- نجفی‌خواه، رقیه، علی جلالی دیزجی، و عصمت مومنی. ۱۳۹۵. امکان‌سنجی ارائه خدمات مرجع مجازی در کتابخانه‌های عمومی استان گیلان. فصل‌نامه مطالعات دانش‌شناسی ۳ (۹): ۸۰-۶۱.
- همتی، عبدالناصر، قدیر مهدوی کلیشمی، و حمید یاری. ۱۳۹۳. امکان‌سنجی ارائه بیمه عمر متصل به سهام در بازار بیمه کشور از دیدگاه خبرگان. دوفصل‌نامه جستارهای اقتصادی ایران ۱۱ (۲۲): ۹۵-۱۱۶.
- یاری، علیرضا. ۱۳۹۴. بررسی پیکره‌ها و ابزارهای پردازش زبان فارسی. تهران: پژوهشگاه ارتباطات و فناوری اطلاعات. پژوهشکده ارتباطات و فناوری اطلاعات.

فهرست منابع لاتین

- AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & F. Oroumchian. 2009. Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems*. Dubai 22(5): 382–387. Elsevier.
- Al-Sulaiti, L., & E. Atwell. 2006. The design of a corpus of contemporary Arabic. *International journal of corpus linguistics* 11(2): 135-171. John Benjamins publishing company.
- Aston, G., & L. Burnard. 1997. *The NBC handbook exploring the British National Corpus with SARA*. Edinburgh University Press.
- Atkins, S., Clear, J., & N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7 (1): 1-16. Oxford University Press (OUP).
- Anderson, J., Beavan, D., & Ch. Kay. 2007. *The Scottish corpus of texts and speech*. In: Beal, J.C., Corrigan, K.P. and Moisl, H.L. (eds.) *Creating and Digitizing Language Corpora*: 17-34. Palgrave, New York, USA.
- Bartsch, S., & S. Evert. 2013. Exploring the Firthian notion of collocation. *Corpus linguistics: Abstract book*: 31-33. Lancaster. UCREL.
- Baus, K., Radimersky, A., Iwanicki, M., & A. Albers. 2014. Feasibility studies in the product development process. *24th CIRP design conference. Procedia CIRP* 21: 473-478. Elsevier.
- Bianchi, F. 2012. *Culture corpora and semantics: methodological issues in using elicited and corpus data for cultural comparison. Chapter 3: corpora and corpus linguistics*. University of Salento.
- Biber, D. 1993. Representativeness in corpus design. *Literary and linguistic computing* 8 (4): 243-257. Oxford University Press (OUP).
- Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & M. Ghayoomi. 2011. Lesson from building a Persian written corpus: Peykare. *Language resources and evolution* 45 (2): 143-164. Springer.
- Burnard, L. 2007. *Reference guide for the British National Corpus. Research technologies service*. Oxford University Computing Services.
- Cavaglia, G. 2002. Measuring corpus homogeneity using a range of measures for inter-document distance. *Proceedings of the Third International Conference on*

Language Resources and Evaluation (LREC'02): 426-431. European Language Resources Association (ELRA). Spain.

- Claude Toriida, M. 2016. Steps for creating specialized corpus and developing an annotated frequency-based vocabulary list. *TESL Canada journal/ revue TESL du Canada* 34 (11): 87-105.
- Dehdarbehbahani, I., Shakery, A., & H. Faili. 2014. Semi-supervised word polarity identification in resource-lean languages. *Neural Networks* 58: 50-59. Elsevier.
- Dehghani, M., Shakery, A., Asadpour, M., & A. Koushkestani. 2013. A learning approach for email conversation thread reconstruction. *Journal of Information Science (JIS)* 39(6): 846-863.
- Durand, J. Gut, U., & K. Gjert. 2014. *The handbook of corpus phonology*. Oxford University Press (OUP).
- Ghayoomi, M., Momtazi, S., & M. Bijankhan. 2013. A study of corpus development for Persian. *International journal on Asian language processing* 20 (1): 17-33.
- Kernaghan, S. 2012. *Technical feasibility assessment*. ISET-international.
- Leech, G. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. adding linguistic annotation*. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.
- McEnery, T., & A. Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press.
- McEnery, T., Xiao, R., & Y. Tono. 2006. *Corpus-based language studies: and advanced resource book*. Routledge. London and New York.
- Megerdooomian, K. 2004. Developing a Persian part-of-speech tagger. *Proceedings of the 1st workshop on Persian language and computer*. 99-105.
- Mukherjee, M., & S. Roy. 2017. Feasibility studies and important aspect of project management. *International journal of advanced engineering and management* 2 (4): 98-100.

- Powers, M. J., Cheney, P. H., & G. Crow. 1990. *Structure system development. Analysis, design, implementation*. Boyd and Fraser publishing company.
- Rasooli, M., Kouhestani, M., & A. Moloodi. 2013. Development of a Persian Syntactic Dependency Treebank. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*: 306-314. Atlanta, USA.
- Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., & A. Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics* 42(5): 950-966.
- Scheid, J. 2013. *An Example of Operational Feasibility*. edited by: Michele McDonough. Bright Hub Inc. <https://www.brighthubpm.com/project-planning/72872-an-example-of-operational-feasibility>
- Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E. et al. 2010. Semi Automatic Development of Farsnet; the Persian Wordnet. *Proceedings of 5th Global WordNet Conference (GWA)*. Mumbai, India.
- Sinclair, J. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. Chapter 1: Corpus and Text — Basic Principles*. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.
- Teubert, W. 2005. My version of Corpus linguistics, *International journal of Corpus linguistics* 10 (1): 1-13.
- Thompson, P. 2004. *Developing Linguistic Corpora: a Guide to Good Practice. Spoken language corpora*. Edited by Martin Wynne .ahds.literature, languages and linguistics. The Oxford Text Archive.
- Wattam, S. M. 2015. *Technological Advances in Corpus Sampling methodology*. Mathematics and Statistics School of Computing and Communications Lancaster University.
- Wayne, M. 2005. *Developing linguistic corpora: a guide to good practice*. Oxbow books. Literary and linguistic computing 22 (1).
- Yang, J., Chen, W., Chen, B., & Y. Jia. 2012. Economic feasibility analysis of a renewable energy project in the rural China. *The 18th Biennial conference of International society for ecological modeling. Procedia environmental sciences*

13: 2280-2283. Elsevier.

- <https://ppp-certification.com/ppp-certification-guide/7-assessing-technical-feasibility>
- <https://ppp-certification.com/ppp-certification-guide/15-assessing-legal-feasibility>
- <https://www.evirtualservices.com/market-feasibility>
- <https://www.evirtualservices.com/resource-feasibility>

Abstract

Corpus design is a complicated and time-consuming process. A well-designed corpus is the one in which the process of feasibility study is done prior to starting the project. Feasibility study tries to illuminate different aspects of the project including the various stages, the cost, human resources, legal rights and the like. It actually helps the project manager(s) to do the project in the best way. Review of literature showed that there exists no study in the field of feasibility study of corpus design. The present study aimed to introduce a comprehensive model for the feasibility study of corpus design. To do so, first the general criteria for corpus design were extracted from related studies. Then, the process of corpus design was explained. Review of literature in the field of feasibility studies led to extracting different aspects and features pertaining to corpus design. Finally, a model was proposed for feasibility study of corpus design. Then, the proposed model was evaluated by specialists using Delphi method. The final comprehensive model has seven aspects (including technical, economical, schedule, legal, operational, resource and marketing) as well as 33 features. To apply the model at Iranian Research Institute for Information Science and Technology (IranDoc), the feasibility study model was evaluated at IranDoc in a panel consisting of specialists. Considering different aspects and features involved in the feasibility study of corpus design, the final model for IranDoc can help the manager in better decision making about corpus design project at IranDoc.

Key words: corpus, feasibility study, model, Delphi



**Iranian Research Institute
for Information Science and Technology (IranDoc)**

**Feasibility study of making a corpus
based on scientific data available at
Iranian Research Institute for
Information Science and Technology
(IranDoc)**

By:

Elham Alayiaboozar

September 2019