زبانِ فارسی و رایانه: برگزیدهٔ مقالات (۲ جلد)، به‌کوششِ: حسین صامتی و محمود بی‌جن‌خان، سازمانِ مطالعه و تدوینِ کتبِ علومِ انسانیِ دانشگاه‌ها (سمت)، ۱۳۸۹، ۱۲۶۶ صفحه.

ما در زمانه‌ای زندگی می‌کنیم که اطلاعات از هر سو به سمتمان روانه است: سپهری از اطلاعات. بخشِ عمدهٔ این اطلاعات در قالبِ زبانِ طبیعی است و ازطریقِ رایانه‌ها به ما می‌رسد. به‌منظورِ دسترسی به این دانشِ ذخیره‌شده و همچنین، ایجادِ میاناهای[1] انسان- ماشین، پردازشِ زبانِ طبیعی، در جامعهٔ اطلاعاتیِ چندزبانهٔ امروزی، نقشِ پررنگ‌تری یافته است. برنامه‌های تصحیحِ دستور و املای کلمات، بازیابیِ اطلاعات در پایگاه‌های داده و ترجمه از زبانی به زبانِ دیگر از پرفروش‌ترین برنامه‌های پردازشیِ زبان‌اند. موفقیت این نرم‌افزارها تخیل فرد را پروار می‌کند، اما باید پذیرفت که این برنامه‌ها همچنان ازنظرِ هوش و کارآمدی، کمبودهای اساسی دارند. در واقع، هدفِ جاه‌طلبانهٔ انسان برای ایجادِ نرم‌افزارِ تولید و درکِ عمیقِ زبانِ طبیعی، که ابزاری قوی برای ترجمهٔ خودکار و ارتباطِ دوسویهٔ ماشین و انسان گردد، همچنان دور از دسترس می‌نماید. به‌نظر می‌رسد برای حلِ مسائلِ همچنان مکتومِ این زمینه‌ها باید به علومِ پایه، و دراین‌مورد، علمِ زبان‌شناسی، توسل جست. مسائلِ زبان‌شناختیِ فارسی، از دیدگاهِ نظری، پیشینه و رشدِ قابلِ توجهی دارند، اما پردازشِ رایانه‌ایِ آن‌ها قدمتِ چندانی ندارد. شاید زمانِ آن رسیده باشد که با بسترسازی و ارائهٔ گسترده‌ترِ محتوای تولیدی، پردازشِ زبانِ فارسی را جدی‌تر بگیریم. کتابِ "زبانِ فارسی و رایانه" نگاهی دوجانبه، زبان‌شناختی و رایانشی، به مسائلِ مربوط به پردازشِ زبانِ فارسی دارد.

۹۴ مقاله‌ای که درقالبِ کتابی باعنوانِ "زبان فارسی و رایانه" در دو جلد و ۱۲۶۶ صفحه تدارک دیده شده‌اند، برگزیده‌ای هستند از مقالاتی که درزمینهٔ زبان و خطِ فارسی در فضای رایانه‌ای تا خردادماهِ سالِ ۱۳۸۶ در هم‌اندیشی‌های مربوط به این‌موضوع، در کنفرانسِ بین‌المللیِ سالانهٔ انجمنِ کامپیوترِ ایران، کنفرانسِ مهندسیِ برقِ ایران، همایشِ زبان‌شناسیِ ایران و کارگاهِ زبانِ فارسی و رایانه، ارائه شده‌اند. به‌جز چهار مقالهٔ انگلیسی (که درانتهای جلدِ دوم چاپ شده‌اند) بقیهٔ مقالات به

1. interface

زبانِ فارسی‌اند. کتاب برحسبِ موضوعِ اصلیِ مقالات به ده بخش تقسیم شده که چهار بخشِ نخست در جلدِ اول و شش بخشِ دیگر در جلدِ دوم ارائه گردیده است. این مقالات طیفِ وسیعی از موضوعاتِ مختلفِ مربوط به گفتار و نوشتارِ فارسی، ازجمله پردازش و بازشناسیِ متن و گفتار، دادگان، ریشه‌یاب و سامانه‌های خبره، را دربر دارد.

بخشِ اولِ کتاب باعنوانِ "پیکره و دادگان" حاویِ دوازده مقاله است. سه مقالهٔ اول به مبحثِ پردازشِ گفتاری درقالبِ تعیینِ مرزِ آوایی (همایون‌پور و احمدی، ۱۳۷۷)، و هجایی (غفاری مقدم و همکاران، ۱۳۷۶)، و واکداری، هجا و واج (همایون‌پور و آزاد، ۱۳۸۰) در گفتارِ فارسی می‌پردازند. ضرورتِ این مرزبندی در تهیهٔ دادگانِ گفتاری آشکار می‌شود.

مقاله‌های شیخ‌زادگان و بی‌جن‌خان (۱۳۸۵) و فاضل و همکاران (۱۳۸۴) و اسلامی و بی‌جن‌خان (۱۳۷۴) به چگونگیِ تهیهٔ دادگانِ گفتاریِ زبانِ فارسی می‌پردازند. مقالهٔ نخست، به معرفی و ذکرِ ویژگی‌های دادگانِ گفتاری می‌پردازد که در پژوهشکدهٔ پردازشِ هوشمندِ علائم طراحی و پیاده‌سازی شده‌اند. دومین مقاله، روندِ طراحی و تهیهٔ دادگانِ تلفنیِ اعدادِ متصلِ فارسی را شرح می‌دهد. این دادگان برای سامانه‌های بازشناسیِ خودکارِ گفتار طراحی شده است. آخرین‌مقاله، گزارشی است از چگونگیِ ایجادِ اولین دادگانِ بزرگِ گفتاریِ زبانِ فارسی (فارس دات)، که شاملِ ۳۸۶ جمله است که ۳۰۰ گویشور از ده منطقهٔ زبانی در گردآوریِ آن نقش داشته‌اند. این مقاله، نیز، به تقطیع و برچسب‌دهیِ آوایی و واجیِ این دادگان می‌پردازد. همچنین، در مقالهٔ پنجم (رضایی و همکاران، ۱۳۸۳) با هدفِ آموزش‌برخط (آنلاین)، تلفظِ واج‌های فارسی به‌صورتِ صوتی- تصویری، به‌همراهِ ارائهٔ پایگاهِ اینترنتیِ آن تدارک دیده شده است. دادگانِ این‌طرح شاملِ سیگنالِ ویدئویی و صوتیِ تلفظِ واج‌ها و گونهٔ رسمیِ تلفظِ الفبای فارسی است.

دو مقالهٔ بعدی به مباحثِ مربوط به واژگان می‌پردازند. در اثرِ نخست (اسلامی و همکاران، ۱۳۸۳)، براساسِ واژگانِ ذهنیِ فارسی‌زبانان و صورت‌بندیِ قواعدِ تصریفِ کلمه در زبانِ فارسی، حدودِ پنجاه‌وپنج‌هزار مدخلِ واژگانی تدارک دیده شده‌است. برنامهٔ واحدسازِ صرفیِ ارائه‌شده با ارجاع به واژگان و قواعدِ تصریف می‌تواند واحدهای زبانی (نوشتار یا گفتار) را به‌لحاظِ صرفی پردازش کند. فامیان و آقاجانی (۱۳۸۵) به‌منظورِ طراحیِ شبکهٔ واژگانیِ صفاتِ زبانِ فارسی، پانزده طبقهٔ اصلی و بیش‌از هفتاد طبقهٔ فرعیِ معنایی ازمقولهٔ صفات ارائه کرده‌اند. رابطِ گرافیکیِ کاربرِ این شبکه با قابلیت‌های جستجو، نمایش و ارسالِ اطلاعات طراحی شده‌است.

مقالهٔ (صفابخش و شایگان، ۱۳۷۴) گزارشی است از تشکیلِ دادگانِ کلمات و حروف به‌منظورِ ایجادِ مجموعه‌ای استاندارد برای آموزش و آزمایشِ همسان و مقایسهٔ عملکردِ الگوریتم‌ها برای بازشناسیِ خودکارِ حروفِ کلماتِ دست‌نویس (نستعلیق). گزارشِ ساختِ نخستین پیکرهٔ چندزبانه برای فارسی (قاسمی‌زاده و همکاران، ۱۳۸۵) برپایهٔ مدلِ MULTEXT به‌معرفیِ ویژگی‌های صرفی-نحویِ فارسی و برچسب‌دهیِ پیکرهٔ متنِ ترجمه‌شدهٔ رمان ۱۹۸۴ به فارسی پرداخته است. چندزبانگیِ این‌پیکره، راه را برای ترجمهٔ بینِ زبان‌ها هموار می‌کند. آخرین مقالهٔ این بخش (حاجی عبدالحسینی، ۱۳۷۶) نیز اولین تلاش برای برچسب‌دهیِ یک پیکرهٔ رایانه‌ایِ فارسی را به‌نمایش می‌گذارد. به‌نظر می‌رسد این مقاله، و تاحدی دومقالهٔ پیشین، بیش‌ازآن‌که به مبحثِ دادگان بپردازند، به روشِ برچسب‌دهیِ آن‌ها پرداخته‌اند.

عنوانِ بخشِ دومِ کتاب، "پردازش متن" است. ازمیانِ دوازده مقالهٔ این بخش، سومین مقاله (شمس‌فرد، ۱۳۸۵) دستاوردها و چالش‌های پردازشِ فارسی را بیان می‌کند. شمس‌فرد ابتدا به بررسیِ پیچیدگی‌ها (ویژگی‌ها)ی پردازشِ زبانِ فارسی، و سپس به معرفیِ فعالیت‌ها و پژوهش‌هایی می‌پردازد که درزمینه‌های پردازشِ واژگانی، ساختواژی، نحوی، معنایی، تولیدی، یادگیری و ترجمهٔ ماشینی، و همچنین تهیهٔ منابعِ زبانی انجام شده‌اند.

سه مقالهٔ اول، دوم و هفتم در صددند تا سامانه‌هایی برای دسته‌بندیِ موضوعیِ متونِ فارسی ارائه کنند. دسته‌بندیِ موضوعی، کاربردهای فراوانی در موتورهای جستجو، پاسخ‌گویی به سؤالات و بازیابیِ اطلاعات دارد. عرب‌سرخی و فیلی (۱۳۸۵) با به‌کاربستنِ مدلِ احتمالاتیِ بیزی Bayesian و ایدهٔ نگهداریِ کلماتِ هم‌آیند برای رسیدن به نتایجِ بهتر، سامانهٔ خود را تهیه کرده‌اند. در مقالهٔ دوم (حاجی‌حسینی و الماس‌گنج، ۱۳۸۵)، اساسِ کار، روشِ تحلیلِ معناییِ پنهان است که در آن نزدیکیِ کلماتی مشخص می‌شود که بیشتر در متن‌های هم‌سنخِ معنایی باهم رخ می‌دهند. این روش قابلیتِ بالایی در جداسازیِ موضوعیِ متون دارد. امامی‌آزادی و الماس‌گنج (۱۳۸۵) روشِ تحلیلِ معناییِ پنهانِ احتمالاتی را (که هافمن برای بازیابیِ اطلاعات ارائه کرده بود) به‌کار برده‌اند. این روش، که زیربنای آماری محکمی دارد، باعثِ شده است شیوه‌های اصولی‌تری در تحلیلِ زبان و متون ایجاد شود. همچنین، با کنترلِ مدل، حینِ آموزش، و حذفِ متغیرهای پنهانِ نامناسب، عملکردِ آن‌را بهبود داده‌اند.

دو مقالهٔ تشکری و میبدی (۱۳۸۲) و بشیری و همکاران (۱۳۸۴) به دو شیوه از طراحیِ نمایه‌سازِ خودکار برای متونِ فارسی می‌پردازند. در هر دو شیوه، ابتدا فهرستی از واژگانِ عمومی تهیه شده است. در مقالهٔ نخست، در روندِ ساختِ نمایه‌سازِ خودکار، پس‌از شناساییِ واژگانِ متن، واژگانِ عمومیِ آن حذف می‌شوند. در مقالهٔ دیگر، الگوریتمِ ریشه‌یابی ارائه‌شده که وندها را مرحله‌به‌مرحله حذف می‌کند تا به ریشهٔ واژه برسد. این الگوریتم دقتِ نمایه‌ساز را بهبود و حجمِ آن‌را کاهش می‌دهد.

شاهمیری و همکاران (۱۳۸۵) درراستای طرح‌های شناساییِ نویسنده از روی نوشته، سامانه‌ای را طراحی کرده‌اند که هدفِ آن شناساییِ اشعارِ شاهنامه است. به‌این‌منظور، پنجاه ویژگیِ اشعارِ فارسی در سه مقولهٔ فیزیکی، مفهومی و آوایی دسته‌بندی شده و سپس، اشعار به دو روشِ پردازشِ آماریِ پارامترها و شبکهٔ عصبیِ مصنوعی تحلیل شده‌اند. همچنین، عظیمی (۱۳۸۵) در مقاله‌اش، با پیش‌نهادنِ دوازده پرسش دربارهٔ ویژگی‌های آوایی، واجی و واژگانی، و بررسیِ آماریِ پاسخ‌ها به تمایزِ شباهت‌ها و تفاوت‌های اشعارِ حافظ، خواجو و سلمان می‌پردازد.

سامانهٔ خلاصه‌سازیِ خودکارِ متونِ فارسی (کریمی و شمس‌فرد، ۱۳۸۵) براساسِ ترکیبی از روش‌های زنجیرهٔ واژگانی و روش‌های مبتنی بر گراف عمل می‌کند. نخستین بخشِ مقاله، به‌معرفیِ خلاصه‌سازی، انواع و نمونه‌های اجرایی‌شدهٔ آن پرداخته است. سپس، با پیش نهادنِ پنج معیار برای امتیازدهیِ جملاتِ متن، جملاتِ بابیشترین امتیاز را به‌عنوانِ خروجی انتخاب می‌کند. مقالهٔ دیگر (صدرموسوی و شمس‌فرد، ۱۳۸۶) برای تشخیصِ روابطِ معنایی میانِ اجزای فعل طراحی شده است. روشِ ارائه شده، مبتنی بر دودسته قوانین، یکی مربوط به هردستهٔ‌معنایی و یکی مربوط به تشخیصِ نقشِ گروه‌های واژگانی است.

سامانه‌های استخراجِ اطلاعات به‌منظورِ پردازشِ اطلاعاتِ مربوط به یک حوزهٔ خاص، در محیط‌هایی که روزانه با حجمِ وسیعی از اطلاعات سروکار دارند، استفاده می‌شود (زبانِ فارسی و رایانه: ۲۸۹). سامانهٔ استخراجِ اطلاعاتِ مِرصاد (رحیمی‌پور و همکاران، ۱۳۸۶) روی اخبار در حوزهٔ نظامی تمرکز دارد. این سامانه، به‌وسیلهٔ الگوهایی متون را پردازش و اطلاعات را استخراج و ذخیره‌سازی می‌نماید. مقالهٔ پایانیِ این بخش، گزارشی است از کارهای انجام‌شده برای ایجادِ پایگاهِ داده‌های زبانِ فارسی در پژوهشگاهِ علومِ انسانی (عاصی، ۱۳۷۴). مقاله به ویژگی‌های دادگان، گسترهٔ زبانی، منابعِ گردآوریِ داده‌ها، ساختارِ زبانیِ پیکره و ساختارِ رایانه‌ای آن، و همچنین کاربری‌های

دادگان (انواعِ جستجو، گزارش‌ها، کاربران و ...) می‌پردازد. البته، این مقاله باید در بخشِ پیشین (پیکره و دادگان) گنجانده می‌شد.

بخشِ ۳: ریشه‌یاب و خطایاب. موتورهای جستجو، سامانه‌های بازیابیِ اطلاعات و ترجمهٔ متون ازجمله زمینه‌هایی هستند که از دستاوردهای ریشه‌یابی برای بهبودِ روش‌هایشان سود جسته‌اند. ازمیانِ ده مقالهٔ ارائه‌شده در این بخش، پنج مقاله به طراحیِ ریشه‌یاب‌های زبانِ فارسی اختصاص دارند. تشکری و میبدی (۱۳۸۰) در مقاله‌شان پس‌از بررسیِ شیوهٔ ساختِ مقوله‌های واژگانی (فعل، اسم، ضمیر) و البته بدونِ درنظرگرفتنِ نقشِ آن‌ها در جمله، الگوریتمی برای ریشه‌یابیِ خودکارِ واژگان طراحی کرده‌اند. مقالهٔ سوم (یوسفان و همکاران، ۱۳۸۵)، پس‌از برشمردنِ دشواری‌های ریشه‌یابیِ زبانِ فارسی، به‌تدارکِ یک فرهنگِ ریشه‌یاب، و درنهایت، روندِ طراحیِ یک شبه‌زبانِ برنامه‌نویسی برای ریشه‌یابیِ افعالِ فارسی پرداخته است. حسامی‌فر و قاسم‌ثانی (۱۳۸۴) درابتدای مقاله‌شان رویکردهای مختلف نسبت‌به مسئلهٔ ریشه‌یابی را مرور کرده‌اند. سپس، بااستفاده از تأثیرِ نقشِ کلمات روی الگوریتم‌های ساختواژهٔ فارسی، از آن برای ریشه‌یابیِ افعالِ فارسی استفاده کرده‌اند. در ریشه‌یابِ ارائه‌شده توسطِ نصیری و همکارانش (۱۳۸۴)، به‌جای استفاده از اطلاعاتِ ساختاریِ زبان، از یک مدلِ احتمالاتی برای ریشه‌یابی استفاده می‌شود. و درنهایت، نهمین مقاله (شاهمیری و همکارانش، ۱۳۸۵) باکمکِ شبکهٔ عصبیِ مصنوعیِ پرسپترون[1] چندلایه به طراحیِ سامانهٔ خودکاری برای تعیینِ ریشهٔ زبانیِ واژگانِ فارسی و عربیِ موجود در زبانِ فارسی می‌پردازد. دقتِ سامانهٔ پیاده‌سازی‌شده تا ۹۲ درصد گزارش شده است.

ازآن‌جاکه حجمِ اطلاعاتیِ وسیعی ازطریقِ دستی یا به‌صورتِ خودکار واردِ رایانه می‌شود و احتمالِ بروزِ خطا در این داده‌ها کم نیست، استفاده از خطایاب‌ها موجبِ کاهشِ هزینه و بالارفتنِ سرعتِ تولیدِ مستنداتِ رایانه‌ای خواهد شد. درّی نوگورانی و صبوریان (۱۳۸۵) در مقالهٔ پنجمِ این بخش، پس‌از معرفیِ روش‌های مختلفِ خطایابی، به طراحیِ خطایابی برای زبانِ فارسی می‌پردازند که در آن پس‌از ریشه‌یابیِ واژه‌های پیوسته، خطاها مشخص و پیشنهادِ لازم به کاربر داده می‌شود. همچنین، سامانهٔ ارائه‌شده در مقالهٔ عرب‌سرخی و همکاران (۱۳۸۵)، بااستفاده از بخش‌های مختلف

1 . multi-layer perception

(واژه‌نامه، الگوریتمِ تشخیصِ خط و الگوریتمِ تصحیحِ خطا)، خطاهای املاییِ زبانِ فارسی را به‌طورِ خودکار تصحیح می‌کند.

موضوعِ مقالهٔ نخستِ این بخش (قدسی و بازرگان، ۱۳۷۴)، ارائهٔ شیوه‌ای نرم‌افزاری برای تولیدِ خودکارِ فونت‌ها به زبانِ متافونت ازراهِ پویشِ تصویر است. ایدهٔ ارائه‌شده دراین‌جا برای بازشناسیِ نوریِ نویسه نیز قابل استفاده است. در مقالهٔ هشتم هم، مرادی و شیری قیداری (۱۳۸۴) به‌دنبالِ رسیدن به چینشِ بهینه‌ای از جایگذاریِ ۳۲ حرفِ الفبای فارسی و همزه برروی صفحه‌کلید هستند. ویژگی چنین بهینه‌سازی، جلوگیری از جابه‌جاییِ زیادِ انگشتان و تایپِ دو حرفِ متوالی با یک انگشت (دست) و تقسیمِ بارِ تایپ بینِ دو دست است.

فولادی و ارومچیان (۱۳۸۵) با بهره‌گیری از پیکره به‌عنوانِ تنها منبعِ یادگیری و با ارائهٔ ساختواژه به‌صورتِ کمینه (تقسیم‌بندی به بن و پسوند)، نحوهٔ استخراجِ ساختواژه را با یادگیریِ بدونِ نظارت ارائه می‌کنند.

دوازده مقالهٔ بخشِ چهارمِ کتاب به ارائهٔ الگوریتم‌هایی برای تشخیصِ نوریِ نویسه (OCR) و تشخیصِ دستخط اختصاص دارد. در مقالهٔ فهیمی و تیمسار (۱۳۷۲) الگوریتمی ارائه شده است که بااستفاده از روشِ ریخت‌شناسی می‌تواند حروفِ تایپ‌شدهٔ فارسی را بازشناسی کند. در مقالهٔ دوم (عزمی و کبیر، ۱۳۷۶)، دو الگوریتمِ مختلف برای بازشناسیِ حروفِ چاپیِ فارسی روی بیست‌نوع فونتِ مختلف آزموده شده و میزانِ بازشناسیِ فونت‌ها توسط این الگوریتم، بالا گزارش شده است. الگوریتمِ بعدی، توسطِ عباسیان و کبیر (۱۳۷۷) و براساسِ ویژگی‌های پویا و ایستای نویسه‌های فارسی تنظیم شده است. در (سلیمانی و همکاران، ۱۳۸۵) الگوریتمی ارائه شده است که با میانگینِ زمانیِ ۰.۰۰۵ ثانیه برای هرحرف به‌محاسبهٔ داده‌های دست‌نوشته می‌پردازد و در تدوینِ آن از ترکیبی از رویکردهای فازی، ساختاری و آماری استفاده شده است. مظفری و همکاران (۱۳۷۵) به طراحی و پیاده‌سازیِ سامانه‌ای برای بازشناسیِ ارقامِ دست‌نویس بااستفاده از روش‌های ساختاری می‌پردازند. در مقالهٔ بعدی (مسروری و همکاران، ۱۳۷۹)، براساسِ استخراجِ ویژگی‌ها از کُلِّ کلمه، روشی پیشنهاد شده است که در آن از پروفایل‌های تصویر استفاده می‌شود. ازآن‌جاکه بازشناسیِ قلم برای بازشناسیِ نوریِ نویسه‌ها اجتناب‌ناپذیر است، نظام‌آبادی‌پور و همکارانش (۱۳۸۲)، روشی برای استخراجِ ویژگی‌های مناسب از تصویرِ متن و بازشناسیِ قلمِ آن ارائه کرده‌اند که نتایجِ آن مطلوب ارزیابی شده است.

دستهٔ دیگرِ مقالاتِ این بخش، باهدفی یکسان، به مباحثِ مربوط به جداسازیِ حروف پرداخته‌اند: مقالهٔ چهارم (نظام‌آبادی‌پور و همکارانش، ۱۳۷۹) باهدفِ آشکارسازیِ نقاطِ جداسازی در متونِ چاپیِ قدیمی به اصلاحِ یک الگوریتمِ قدیمی پرداخته است. صفابخش و ادیبی (۱۳۸۱) الگوریتم‌هایی را برای مراحلِ پیش‌پردازش، تقطیع و یافتنِ ترتیبِ راست‌به‌چپِ قطعه‌ها برای کلماتِ دست‌نویسِ فارسی ارائه کرده‌اند که نتایجِ عملکردشان درحدِ مطلوبی گزارش شده است. جداسازیِ تصویریِ کلماتِ فارسی و لاتین بااستفاده‌از یادگیریِ استقراییِ قواعد (شش ویژگیِ برگرفته از تصویر و تدوینِ سیزده قاعدهٔ کلی) موضوعِ مقالهٔ صدوقی یزدی و همکارانش (۱۳۸۱) است. "زنجیرِ فعال" روشی است که از آن برای یافتنِ مرزِ چیزهای موجود در تصویر استفاده می‌شود. گرایلو و کبیر (۱۳۸۵) به بررسی و ارزیابیِ این روش و چگونگی بهبودِ عملکردِ آن در اصلاحِ گسستگی‌های ناخواسته در متونِ چاپیِ فارسی می‌پردازند.

درنهایت، مقالهٔ شهابی‌نژاد و رحمتی (۱۳۸۵) بامروری بر مطالعاتِ تعیینِ هویتِ مبتنی‌بر دستخط، روشی برای تعیین و تأییدِ نویسنده به‌صورتِ برون‌خط (مبتنی بر متن) ارائه می‌کند.

بخشِ مربوط به ترجمهٔ ماشینی نیز دوازده مقاله دربر دارد. در ابتدای این بخش، فضلی و فهیمی (۱۳۷۷) به بررسیِ چند شیوهٔ بازنماییِ معنایی، ازجمله گراف‌های مفهومی پرداخته و نشان می‌دهند که می‌توان همین شیوه را گسترش داد تا دامنهٔ وسیعی از جمله‌های فارسی را بپوشاند. سه مقالهٔ بعدی (قاسم‌ثانی و سلیمانی، ۱۳۷۸؛ فیلی و قاسم‌ثانی، ۱۳۸۳ الف؛ فیلی و قاسم‌ثانی، ۱۳۸۳ ب)، به‌ترتیب با اساس‌قراردادنِ روش‌های تجزیهٔ چارت، دستورِ درخت افزایشیِ هم‌زمان و دستورِ یکسان‌سازی (مدلِ توسعه‌یافتهٔ دستورِ مستقل‌ازمتن) سامانه‌های مترجمیِ مختلفی را معرفی می‌کنند.

سه مقالهٔ بعد مربوط‌اند به ساختِ دستورهای رایانه‌ای: سجادی و عبدالله‌زادهٔ بارفروش (۱۳۸۵) به‌بررسیِ نحوهٔ عملکردِ دستورِ پیوندی (دستورِ کلمه) برای زبانِ فارسی، عرب‌سرخی و همکاران (۱۳۸۵) به ارائهٔ ایدهٔ تولیدِ قواعدِ دستوری بااستفاده از الگوریتم‌های ژنتیک، و شمس‌فرد و همکاران (۱۳۸۶) به ساختِ برخی ابزارها و پیمانه‌های پایه‌ایِ پردازشِ زبانِ فارسی می‌پردازند. همچنین، مگردومیان (۱۳۸۳، ۱۳۸۵) در دو مقالهٔ پایانیِ این بخش به‌ترتیب به تحلیلِ معناییِ جزءِ فعلیِ افعالِ مرکب و شیوهٔ ترجمهٔ این افعال، و معرفیِ یک تحلیل‌گرِ صرفی برای زبانِ مورداستفاده در وبلاگ‌های فارسی بااستفاده از فن‌آوریِ ارائه‌شده در مرکزِ زیراکس می‌پردازد.

درنهایت، سه مقالهٔ پایانیِ این بخش، به سامانه‌های برچسب‌زنِ متنِ فارسی پرداخته‌اند. برچسب‌دهیِ متن در زمینه‌های بازشناسیِ گفتار، بازیابیِ اطلاعات، رفعِ ابهامِ معنیِ کلمات و تجزیهٔ جمله‌ها کاربرد دارد. ثابتی و همکارانش (۱۳۸۶) به دو روشِ احتمالی و برمبنای تغییرِ سامانهٔ برچسب‌زن، متنِ فارسی را پیاده‌سازی کرده‌اند. مقالهٔ کیخا و همکاران (۱۳۸۶) گزارشی است از کاربردِ درخت‌های تصمیم (از روش‌های یادگیریِ ماشین) در برچسب‌دهیِ متن و تهیهٔ برچسب‌زن MFT به‌صورتِ مستقل‌ازمتن. در مقالهٔ مگردومیان (۱۳۸۳) نیز به چالش‌های اساسی در تدوینِ برچسب‌زنِ فارسی و راه‌هایی برای حلِ آن‌ها پرداخته شده است.

سه‌بخشِ بعدیِ کتاب، به مباحثِ پردازشِ گفتار اختصاص دارند. در بخشِ ششم، دوازده مقاله باموضوعِ تبدیلِ متن به گفتار ارائه شده است. نخستین مقاله (ابوطالبی و تبیانی، ۱۳۷۸) بامقدمه‌ای دربارهٔ روش‌های بازسازیِ گفتار آغاز می‌شود و باانتخابِ روشِ اتصالِ هجاها و بررسی‌های دیگر، سامانه‌ای برای تبدیلِ متنِ فارسی به گفتار را پیاده‌سازی می‌کند. مقالهٔ دوم (شیخ‌زادهٔ نجار و همکاران، ۱۳۷۸) هم تحقیقی است برای تطبیق و پیاده‌سازیِ گفتارسازِ فارسی برمبنای گفتارسازِ کلات (Klatt). مقالهٔ اسلامی و همکارانش (۱۳۸۳) گزارشی است از بستهٔ نرم‌افزاریِ "گویا" که نوشتارِ فارسی را به گفتار تبدیل می‌کند. همچنین، همایون‌پور و نم‌نبات (۱۳۸۳) برپایهٔ روشِ بازسازیِ پیوندی مبتنی‌بر انتخابِ واحد، سامانهٔ بازسازیِ گفتارِ "فارس بیان" را ارائه کرده‌اند. در مقالهٔ انگلیسیِ شیخان و همکاران (۱۳۸۵) نیز سامانهٔ گفتارسازِ فارسیِ دیگری ارائه شده که از سه زیربخش تشکیل شده است: پردازش‌گرِ زبان، تولیدگرِ نوای گفتار و تحلیل‌گرِ گفتار.

سامانه‌های تبدیلِ متن به گفتار از دو بخش تشکیل می‌شوند: تبدیلِ متن به زنجیرهٔ واج‌های تشکیل‌دهندهٔ آن، و تبدیلِ زنجیرهٔ واج‌ها به گفتار. مقاله‌های بعدی عمدتاً روی بخشِ نخستِ این سامانه‌ها تمرکز دارند. همایون‌پور و حسینی‌نژاد (۱۳۷۸) برای تبدیلِ متن به واج‌های نظیرش از شبکهٔ عصبیِ پرسپترونِ چندلایه سود جسته‌اند. نم‌نبات و همکاران (۱۳۸۵) بااستفاده از همین شبکهٔ عصبی به مدل‌سازیِ دیرشِ واج پرداخته‌اند. مدل‌کردنِ دیرشِ واج بااستفاده از روشِ مارس نیز در مقالهٔ دیگری (بحرینی و همایون‌پور، ۱۳۸۵) بررسی شده است. مقالهٔ نم‌نبات و کوچاری (۱۳۸۶) درنگِ میانِ کلمات را بااستفاده از درختِ دسته‌بندی مدل‌سازی کرده است. و ابوطالبی و فغانی (۱۳۸۵) درراستای میزانِ فهم‌پذیریِ سیگنالِ گفتار، به ارائهٔ دادگانِ فارسی برای آزمونِ قافیهٔ اصلاح‌شده می‌پردازند.

دو مقاله دیگر، یکی (معطر و همکاران، ۱۳۸۳) است که بااستفاده از الگوی عبارات، به‌طورِ مبسوط به نرمال‌سازی در زبانِ فارسی می‌پردازد. نرمال‌سازی، تبدیلِ صورت‌های‌ناموجوددرفرهنگ‌ها به‌صورتی است که معادلِ واجنگاریِ آن‌ها در فرهنگ موجود باشد. مقالهٔ دیگر، نوشته بی‌جن‌خان و مرادزاده (۱۳۸۳)، به‌مسئلهٔ هم‌نگاره‌های خطِ فارسی پرداخته‌است. توصیفاتِ ارائه‌شده در حلِ مسئلهٔ هم‌نگاره‌های فارسی در سامانه‌هایی چون تبدیلِ متن به گفتار و ترجمهٔ ماشینی بسیار کاربرد دارد.

دربخشِ بازشناسیِ گفتار هم دوازده‌مقاله ارائه شده است. دراین‌میان، پنج مقاله (شیخان و همکاران، ۱۳۷۳؛ پرویزی و احدی، ۱۳۸۰؛ الماس‌گنج و همکاران، ۱۳۸۰؛ صامتی و همکاران، ۱۳۸۳؛ الماس‌گنج و همکاران، ۱۳۸۳) به‌معرفیِ سامانه‌های بازشناسیِ گفتارِ فارسی می‌پردازند. در مقالهٔ نخست، از شبکه‌های عصبیِ مصنوعی و روش‌های هوشِ مصنوعی استفاده شده است، در دومی، با دایرهٔ کلماتِ متوسط، گفتارِ پیوستهٔ فارسی بازشناسی می‌شود. دو مقالهٔ الماس‌گنج و همکارانش سامانه‌های شنوا ۱و ۲ را معرفی می‌کنند. و مقالهٔ صامتی و همکارانش گزارشی است از سامانه‌ای که به‌صورتِ مستقل‌ازگوینده عمل می‌کند.

مقالاتِ دیگرِ این بخش، هریک به بحث‌وبررسیِ بخشی از مسائلِ مربوط به سامانه‌های بازشناسیِ گفتار می‌پردازند. بررسی و ایجادِ ساختار برای همگونیِ واج‌های مجاورِ زبانِ فارسی (صامتی و همکاران، ۱۳۷۶)، راه‌هایی برای رفعِ مشکلاتِ ناشی‌از ماهیتِ سیگنال (الماس‌گنج و همکاران، ۱۳۷۷)، روشی برای حلِ مسئلهٔ وابستگی‌به‌بافت در تلفظِ کلمات (غلامپور و همکاران، ۱۳۷۹)، بازشناسیِ محلِ هجای تکیه‌برِ کلمات (دهقانِ دهنوی و همکاران، ۱۳۸۰)، و استفاده از ویژگیِ نوایی و میزانِ تکیهٔ هجاهای فارسی در فرایندِ بازشناسیِ گفتارِ پیوستهٔ فارسی (الماس‌گنج، ۱۳۸۱) ازموضوعاتِ مختلفِ موردِ بحث است.

باتوجه‌به‌آن‌که سامانهٔ بازشناسی در انسان ساختاری سلسله‌مراتبی و دوسویه دارد، در مقالهٔ یزدچی و همکارانش (۱۳۸۵) کارآییِ شبکه‌های عصبیِ دوسویه در بازشناسیِ گفتار بررسی شده‌است؛ و درانتها، مقالهٔ ویسی‌پور و همکاران (۱۳۸۵) برای مواجهه باکلماتِ خارج از واژگان (ناآشنا برای سامانهٔ بازشناسی) از معیارِ اطمینان استفاده می‌کند، که مبتنی است بر امتیازدهیِ کلمات برای اطمینان از صحتِ بازشناسیِ آن‌ها.

ازمهم‌ترین راه‌های بهبودِ عملکرد، افزایشِ دقت و کاهشِ خطای سامانه‌های بازشناسیِ گفتار، استفاده از مدل‌های زبانی (اطلاعاتِ آماری، دستوری و ...) است، و بخش هشتم هم به "مدل‌های

زبانی در بازشناسیِ گفتار" اختصاص دارد. مراحلِ طراحی و پیاده‌سازیِ یک رمزگشای واج به‌کلمه (باباعلی و همکاران، ۱۳۸۳)، استفاده از تلفیقی‌از مدل‌های دستوریِ احتمالاتی به‌عنوانِ مدلِ زبانی (ممتازی و همکاران، ۱۳۸۴)، استفاده از مدلِ دستورِ ساختِ گروهیِ تعمیم‌یافته برای ساختِ مدلِ زبانی (حافظی و همکاران، ۱۳۸۵)، استفاده از پیکرهٔ متنیِ زبانِ فارسی برای استخراجِ مدلِ زبانی (بحرانی و همکاران، ۱۳۸۵)، و استفاده از روشِ تحلیلِ معناییِ پنهانِ احتمالاتی به‌منظورِ پرداختن به بُعدِ معناییِ زبان (امامی آزادی و همکاران، ۱۳۸۵)، موضوعاتِ مقالاتِ مطرح‌شده در این بخش‌اند.

شش مقالهٔ بخشِ نهم، سامانه‌های خبره‌ای را برای زبانِ فارسی ارائه کرده‌اند. نخستین مقاله، معرفیِ سامانهٔ "پرس‌وجوی فارسی" است (سلطانی خسروشاهی و لوکس، ۱۳۷۹): یک رابطِ زبانِ فارسی برای دادگانِ رابطه‌ای، که به دادگانِ خاصی هم وابسته نیست. فهیمی و شمس‌فرد (۱۳۷۴) سامانهٔ "دنا" را بااستفاده از روشِ وابستگی‌های مفهومی برای درکِ متنِ فارسی تدوین کرده‌اند. درکِ متن عبارت‌استاز تبدیلِ ورودیِ زبانِ طبیعی به یک بازنماییِ داخلیِ قابلِ پردازش توسطِ ماشین. "هستی" سامانه‌ای جهتِ استخراجِ دانشِ واژگانی و مفهومی از متونِ نوشتاریِ زبانِ فارسی و ساختِ واژگان و هستان‌شناسی[1] براساسِ آن‌ها است. شمس‌فرد (۱۳۸۳) به‌معرفیِ این سامانه می‌پردازد. "دانا" (داودآبادی و پالهنگ، ۱۳۸۴) هم سامانه‌ای است که پس‌از دریافتِ یک جملهٔ فارسی و تحلیلِ نحوی و معناییِ جمله، ساختارِ ویژگی و قابِ حالتِ جملهٔ دریافتی را می‌سازد و به فرمانِ کاربر پاسخ می‌دهد. سامانه‌های پرسش‌وپاسخ شاخه‌ای از تحقیقاتِ گروهِ بازیابیِ اطلاعات‌اند که به‌ویژه برای کاربرد در موتورهای جستجو به‌وجود آمده‌اند. در این سامانه‌ها سعی بر آن است که سؤالِ کاربر باجوابی کوتاه، دقیق و کامل پاسخ داده شود. مقالهٔ شمس‌فرد و همکاران (۱۳۸۵) بررسیِ معماری و عملکردِ یک سامانهٔ پرسش‌وپاسخ به‌زبانِ فارسی است. رجبی و همکاران (۱۳۸۵) به‌معرفیِ روش‌هایی جهتِ تبدیلِ جمله‌های زبانِ فارسی به بازنماییِ منطقی (درکِ متن) و همچنین تبدیلِ عبارت‌های منطقی به جمله‌های فارسی (تولیدِ متن) می‌پردازند. در این مقاله، برای روندِ درک، از روشِ تحلیلِ معناییِ مبتنی‌بر نحو و برای تولید از نظریهٔ معنا- متن استفاده شده است.

بخشِ دهم باعنوانِ "سایرِ موضوعات" مقالهٔ شیخ‌زادگان و همکاران (۱۳۷۴)، "بررسیِ درجهٔ اهمیتِ واج‌های زبانِ فارسیِ گفتاری و تعیینِ ترتیبِ آن‌ها ازنظرِ شناساییِ گوینده" را دربر دارد، و

1. ontology

درواقع، مطالعهٔ درجهٔ اهمیتِ واج‌ها ازنظرِ قدرتِ تفکیک و تمایزِ گویندگانشان است و ازنتایجِ آن در سامانه‌های شناساییِ گوینده استفاده می‌شود.

در این کتاب به بسیاری از موضوعاتِ مطرح در پردازشِ رایانه‌ایِ زبانِ فارسی پرداخته شده، و از طیفِ مقالاتِ ارائه‌شده به‌خوبی نمایان است که تکمیلِ پروژه‌های موجود و پرداختن به جنبه‌های کمترکارشده نیاز به کارهای اساسی در آینده دارد. تدوینِ این‌مجموعه به‌همتِ دو استادِ ساعی، که خود درزمینه‌های مختلفِ دادگان و پردازشِ گفتار سال‌ها پژوهش کرده‌اند، بسیار ارزنده است، به‌ویژه‌آن‌که این مجموعه نیازِ جامعهٔ زبان‌شناسانِ رایانه‌ای به دردست‌داشتنِ منسجمِ دست‌آوردهای پردازشِ زبانِ فارسی را برآورده می‌سازد. تقسیم‌بندیِ بخش‌های مختلفِ کتاب، موضوعاتِ عمدهٔ مطرح در زبان‌شناسیِ محاسباتی را نشان می‌دهد، هرچند دربرخی موضوعات، تحقیقاتِ چندانی درزمینهٔ زبانِ فارسی دیده نمی‌شود.

محدودشدنِ کتاب به کنفرانس‌ها و همایش‌های نام‌برده‌شده، ازنظرِ محدودیتِ امکاناتی، که همیشه وجود داشته و دارد، و همچنین بالارفتنِ حجمِ کتاب، قابلِ درک است، اما ازنظرِ محدودیت در انتخابِ مقالاتِ اصیل و ارزنده درزمینهٔ پردازشِ زبانِ فارسی، که اتفاقاً در این همایش‌ها ارائه نشده‌اند، خلأ بزرگی را ایجاد می‌کند. به‌ویژه‌آن‌که این‌کتاب می‌تواند منبعِ ارزنده‌ای برای دانشجویانی شود که بنابر نیازِ روبه‌گسترشِ تحقیقات درزمینهٔ پردازشِ زبانِ فارسی، به‌سویِ این زمینهٔ تحقیقاتی گرایش می‌یابند. ازطرفِ دیگر، به‌نظر می‌رسد آوردنِ تاریخچهٔ مختصری دربارهٔ هرموضوع و اهدافِ پیشِ روی آن برای زبانِ فارسی، خواننده را در درکِ موقعیتِ زبانِ فارسی در رایاسپهر یاری می‌کرد.

از بهترین نکاتِ این کتاب، تدوینِ واژه‌نامهٔ برابرنهادهای اصطلاحاتِ تخصصی و یکسان‌سازیِ این اصطلاحات (تاحدِ ممکن) در متنِ مقالات است. این‌کار از سردرگمی یا بدفهمیِ خوانندگان جلوگیری کرده و به کلّیتِ کتاب نیز نظم و هماهنگی بخشیده است. ازدیگرسو، نداشتنِ نمایه را شاید بتوان بزرگ‌ترین کمبودِ این‌مجموعه دانست. ازآن‌جاکه این دو جلد به‌عنوانِ مرجعی برای تحقیقاتِ رایانه‌ای روی زبانِ فارسی عمل خواهند کرد، بهتر بود دست‌کم نمایه‌ای موضوعی برای آن تدوین می‌شد.

درپایان، امید است همان‌طور که ویراستارانِ کتاب نیز ابراز کرده‌اند، در آینده روندِ تدوینِ مجموعه‌مقالاتِ اصیل در این حوزه ادامه یابد.

صدیقه مرادی

منابع

عاصی، مصطفی (۱۳۸۵). "زبان فارسی در رایاسپهر: جایگاه زبان فارسی در جهان نوین فن‌آوری اطلاعات". *نامهٔ فرهنگستان*. پاییز ۱۳۸۵، شماره ۳۱. ۷۰-۵۹.

Bird, Steven and Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Bolshakov. Igor A. and Alexander Gelbukh (2004). *Computational Linguistics: Models, Resources, Applications*. Instituto Politecnico Nacional: Mexico.