بررسی مزایا و معایب استفاده از منابع اینترنتی به‌عنوان بخشی از پیکرهٔ فرهنگ‌نویسی

سارا شریف‌پور (عضو هیئت‌علمی فرهنگستان زبان و ادب فارسی)

چکیده: در این مقاله استفاده از شاهدهای اینترنتی در فرایند فرهنگ‌نویسی بررسی شده‌است و به این منظور شاهدهایی از فرهنگ جامع زبان فارسی ارائه شده که نشان می‌دهد مؤلفان این فرهنگ در کنار بهره‌مندی از پیکرهٔ زبانی عظیم خود، به‌دلیل نزدیکی شاهدهای اینترنتی به کاربرد روزمره و طبیعی زبان، نیازمند استفاده از این شاهدها نیز هستند. با توجه به این نیاز، در مقالهٔ پیش رو با استفاده از تحلیل ترپ و فوئرتس ـ اولیورا (۲۰۱۶)، مزایا و حساسیت‌های استفاده از شاهدهای اینترنتی معرفی شده و به بررسی نقش فعال مؤلفان فرهنگ در استفادهٔ بهینه و مؤثر از داده‌های اینترنتی پرداخته شده‌است. در ادامه، براساس این تحلیل، رویکردهای اتخاذشده در فرهنگ جامع زبان فارسی برای استفاده از شاهدهای اینترنتی بررسی شده و با به‌دست دادن نمونه‌هایی از مدخل‌های این فرهنگ، آن‌ها آسیب‌شناسی و تحلیل شده‌است.

کلیدواژه‌ها: پیکرهٔ فرهنگ‌نویسی، پیکرهٔ اینترنتی، شاهدهای زبانی اینترنتی، استخراج شاهدهای اینترنتی

۱ـ مقدمه

پایه و اساس تألیف هر فرهنگی را پیکرهٔ آن تشکیل می‌دهد. پیکره مجموعه‌ای از داده‌های نوشتاری یا گفتاری است که فرهنگ‌نویسان از آن‌ها برای استخراج اطلاعات گوناگون و متعددی دربارهٔ یک مدخل واژگانی، مانند اطلاعات معنایی، دستوری، ریشه‌شناسی، کاربردشناختی و غیره بهره می‌برند. حجم اطلاعاتی که برای هر مدخل واژگانی در یک فرهنگ ارائه می‌شود، بستگی به محدودهٔ کاربری تعریف‌شده و کاربران مفروض آن دارد.

عموماً پیکره‌های زبانی شـامل منابع زیر هسـتند: منابع نوشـتاری که می‌تواند از کهن‌ترین نوشته‌ها تا زمان تألیف فرهنگ، شامل کتاب‌ها و مقاله‌هایی در موضوعات گوناگون، نامه‌ها، سـفرنـامه‌ها و غیره را دربر بگیرد. منابع گفتاری که حاوی گفت‌وگوهای روزمرهٔ اهل زبان، به‌ویژه برای ثبت تلفظ صـحیح یا امروزی واژه‌ها، اسـت و منابع حاصـل از پژوهش‌های میدانی که به‌ویژه درمورد اصطلاحات مربوط به حرفه‌ها و پیشه‌ها کاربرد دارد (خطیبی ۱۳۸۶، ص ۱۴).

خطیبی (۱۳۸۶) در مقالهٔ خود، به معرفی و بررسـی نقش پیکره و کهن‌ترین فرهنگ‌های فارسـی از حدود قرن ۵ هجری تا دوران معاصـر پرداخته‌است. آنچه از این پژوهش دریافت می‌شـود این اسـت که فرهنگ‌ها عموماً متشـکل از منابع نوشتاری، که سهم قابل توجهی از آن‌ها به پیکرهٔ متون منظوم و منثور زبان فارسـی اختصـاص داشـته‌است و نیز در برخی موارد منابع حاصـل از پژوهش‌های میدانی بوده‌است. این مسئله شامل فرهنگ‌های تألیف‌شده در دورهٔ معاصر هم می‌شود.

در میان فرهنگ‌های معاصـر فارسـی، فرهنگ جامع زبان فارسی توانسته‌است به اقتضای گسترش فضای مجازی و دنیای اینترنت در سراسر جهان و همسو با سیاست سایر فرهنگ‌ها در کشـورهای دیگر، از این فضـا به‌عنوان منبع زبانی جدید و غنی بهره بَرَد که امری بی‌سابقه در تاریخچهٔ فرهنگ‌نویسی در زبان فارسی است.

۲ـ منابع اینترنتی به‌عنوان بخشی از پیکره

حضـور رایانه در مراحل گوناگون فرهنگ‌نویسـی را شـاید بتوان همانند انقلابی در این حوزه درنظر گرفـت. بـارزترین نمود این حضـور، جنبـهٔ فنی و وجـه کـاربرد ابزاری رایـانـه در فرهنگ‌نویسـی است که امتیاز اصلی آن، توانایی در سامان‌دهی سریع داده‌ها، حفظ انسجام و ترتیـب آن‌هـا و بـازسـازی آن‌هـا در قالب زیرمجموعه‌ها (پیوندها و ارجاعات و غیره) با اسـتفاده از نرم‌افزارهای فرهنگ‌نویسـی اسـت. همچنین گردآوری همایندهای موجود در بـافـت‌هـا و تا حدودی تعیین طبقات و الگوهای نحوی و در کل امکان پرداختن به جزئیات بیشتر نیز از مواردی است که رایانه می‌تواند به فرآیند فرهنگ‌نویسی کمک نماید. مورد دیگر نقش پررنگ و برجستهٔ رایانه‌ها در حوزهٔ انتشـار فرهنگ‌هاسـت که موجب صـرفه‌جویی در منابع مادی و افزایش چشـمگیر سـرعت و زمان مورد نیاز برای چاپ یا ویرایش فرهنگ‌ها می‌شود (**Zgusta** 1989).

اما ورود رایانه به حوزهٔ معنی به اندازهٔ جنبه‌های فنی و ابزاری سریع و قدرتمند نبوده‌است. تشـخیص معنای ناشـناختهٔ واحدهای واژگانی بر پایهٔ انبوه داده‌های زبانی، با توجه به حشـو موجود در بافت‌های طبیعی زبان و مسائل بسط استعاری و مجازی معنی امری دشوار است کـه هنوز رایـانه قادر به انجام دقیق آن نیسـت. به عبارت دیگر، توصـیف و تحلیل معنی و طبقه‌بندی بافت‌ها و شـاهدهای مورد نیاز در فرهنگ‌نویسی یا به تعبیر دقیق‌تر، بخش خلاق بهره‌گیری از داده‌ها، عمدتاً همچنان به‌عهدهٔ تعریف‌نگاران و ویراسـتاران فرهنگ اسـت. با وجود این، نباید از نقش تأثیرگذار و البته کمّی رایانه‌ها در حوزهٔ معنی و مشخصاً در گردآوری پیکره‌ها غافل شـد که به‌صـورت غیرمسـتقیم و با فراهم کردن امکان دسـترسی به انبوهی از داده‌ها، کمک چشمگیری به مؤلفان فرهنگ در تشخیص و تفکیک معنی می‌کند.

فوئرتس ـ الیورا (Fuertes-Olivera 2012, p. 51) معتقد است که پیکرهٔ فرهنگ‌نویسی را (یعنی پیکره‌ای کـه بتوان از آن در تألیف فرهنگ بهره برد) می‌توان به این صـورت تعریف کرد: هر مجموعه‌ای از متون (نوشـتاری یا گفتاری) که فرهنگ‌نویسـان از آن برای تکمیل سـاختار فرهنـگ و تـألیف آن بهره می‌برند. درنتیجه، متون موجود در فضـای اینترنت را هم به همین تعبیر می‌توان مجموعـه‌ای از متون در نظر گرفت که بر اسـاس این تعریف می‌تواند به‌عنوان پیکرهٔ فرهنگ‌نویسـی به‌کار گرفته شـود. این همان نظری اسـت که کیلگاریف و گرفنسـتت (Kilgarriff and Grefenstette 2003, p. 334) نیز آن را تأیید کرده‌اند و به این پرسـش که آیا اینترنت را می‌توان یک پیکره در نظر گرفت، به‌روشنی پاسخ مثبت داده‌اند.

برای استفاده از منابع اینترنتی به‌عنوان بخشی از پیکرهٔ فرهنگ‌نویسی دو روش می‌توان در نظر گرفت، یکی ایجاد یک پیکره براساس متون موجود در اینترنت و دیگری استفادهٔ مستقیم از اینترنت به‌عنوان نوعی پیکره.

در کشـور ما، برخلاف دیگر کشورها، استفاده از اینترنت به‌عنوان پیکرهٔ فرهنگ‌نویسی یا بخشـی از آن هنوز چندان به‌کار گرفته نشـده یا حتی تأیید نشـده‌است. ازاین‌رو، در این رابطه می‌توان فرهنگ جامع زبان فارسی را نمونه‌ای پیشرو در کشور ایران در نظر گرفت که علاوه‌بر پیکرهٔ زبانی عظیم خود که مبتنی اسـت بر متون نظم و نثر از اولین آثار مکتوب تا کتاب‌ها، رساله‌ها، مقالات و روزنامه‌های سال‌های اخیر، از منابع اینترنت نیز در فرآیند فرهنگ‌نویسی بهره برده‌اسـت. این ویژگی وجه ممیز بسـیار ارزشـمندی اسـت که این فرهنگ را از سـایر فرهنگ‌های معاصـر متمایز می‌کند. از میان فرهنگ‌های فارسـی معاصـر، شاید تنها فرهنگ بزرگ سـخن این امکان را داشـته‌اسـت که از منابع اینترنتی هم به‌عنوان بخشی از پیکره بهره ببرد که البته این اتفاق نیفتاده و مقایسهٔ آن با فرهنگ جامع زبان فارسی به‌روشنی گویای تأثیر

آشـکار منابع متنی اینترنت در غنای بیشـتر امکانات فرهنگ‌نویسـی به‌ویژه درج معنی‌های بیشـتر، مواجهه با صـورت‌های املایی، گفتاری، عوامانه و جز آن‌ها و نیز دسـترسی به انواع تفاوت‌های سـبکی اسـت. مدخل‌های زیر به‌طور تصـافی از جلد اول فرهنگ جامع زبان فارسـی انتخاب شـده‌اند. اطلاعاتی که برای هر مدخل در زیر برشـمرده شـده‌است، همگی مواردی هستند که تنها با تکیه بر شاهدهای اینترنتی به‌دست آمده و در فرهنگ درج شده‌اند. مشـابه این اطلاعات در هیچ‌یک از فرهنگ‌های فارسـی موجود، حتی فرهنگ بزرگ سـخن، که معنی‌ها و شاهدهای معاصر آن پررنگ است، وجود ندارد:

***فرهنگ جامع زبان فارسی* ـ جلد ۱**

آره و آجرپاره (شبه‌جمله) (خودمانی) (برای تحقیر و توهین) (مجاز) در پاسخ به جواب «آره» کسی و برای اعتراض به آن همراه با عصـبانیت و سـرزنش به‌کار می‌رود.

آره[۳] معنی ۹: با لحن پرسـشـی، برای نشـان دادن تعجب دربرابر گفتهٔ دیگری به‌کار می‌رود.

آزمایش آزمایش دادن: به آزمایشگاه مراجعه کردن و خود را تحت آزمایش پزشکی قرار دادن.

آزمایشـی دو. (قیـد) (درمورد انجام گرفتن امری) بـه‌طور موقـت تـا کیفیـت و کارایی آن برای هدفی مشخص، مورد ارزیابی قرار گیرد.

آژانسـی یـک. (صـفت) مربوط به آژانس. ... دو. (اسم) کسی که در آژانس کار می‌کند.

آس[۴] دو. (صفت) (مجاز) بسیار عالی و بی‌نظیر.

آقا[۱] **آقای گل**: (ورزش) (معمولاً در فوتبال) بازیکنی کـه در مسـابقـه‌هـای تیم خود بـا تیم‌هـای دیگر، بیشترین تعداد گل را به بازیکنان تیم‌های رقیب زده باشد.

آک[۲] صورتِ کوتاهِ آکبند (ص ۸۳۰).

آنتایم یک. (صـفت) (کسـی یا چیزی) که از نظر برنامه‌ریزی و زمان‌بندی، منظم و سروقت باشد.

چنان‌که دیده می‌شـود، دسـترسـی به داده‌های اینترنتی باعث شده‌است که در هر مدخل اطلاعات و جزئیات بیشـتری ارائه شـود که به‌ویژه از نظر تعداد برش‌های معنایی قابل تأمل اسـت. این نکته را نیز باید در نظر داشت که نمونه‌های یادشده «تنها شامل مواردی بودند که درج یـک برش معنـایی جدید یا ایجاد یک زیرمدخل و اطلاعات دیگری مانند آن، صـرفاً مبتنی بر منابع اینترنتی بوده‌اند و از آوردن نمونه برای مواردی که در کنار شـاهدهای پیکره از شـاهدهای اینترنتی هم اسـتفاده شـده، صـرف‌نظر شده‌است». ناگفته پیداست که در چنین مواردی نیز بهره‌گیری از منابع متنی اینترنت هرچند به درج اطلاعی جدید منجر نشده‌است، اما بی‌تردید بر غنای شاهدها و اطلاعات ارائه‌شده در فرهنگ افزوده‌است.

امـا آیـا اسـتفاده از منابع متنی اینترنت، در هر شـکل و با هر کیفیتی مطلوب اسـت؟ فرهنـگ‌نویسـان در این باره چه قاعده‌هایی دارند؟ و اینکه مؤلفان فرهنگ جامع در این باره چگونه عمل کرده‌اند؟

۳ـ بررسی مزایا و معایب استفاده از اینترنت

برای پاسخ دادن به پرسش‌های مطرح‌شده، در این بخش مزایا و معایب و حساسیت‌های استفاده از منابع اینترنتی به‌عنوان بخشی از پیکرهٔ فرهنگ بررسی می‌شود تا امکان بررسی عملکرد فرهنگ جامع زبان فارسی نیز فراهم شود.

ترپ و فوئرتس ـ اولیورا (**Tarp** and **Fuertes-Olivera** 2016, p. 277) در پژوهش خود نکته‌ها و ملاک‌هایی را برای ارزیابی نتایج استفاده از منابع متنی اینترنت در فرهنگ‌ها معرفی کرده‌اند. ما نیز در مقالهٔ پیش رو از آن نکته‌ها و ملاک‌ها برای بررسی و ارزیابی فرهنگ جامع زبان فارسی از این منظر بهره برده‌ایم:

فرهنگ‌نویسان می‌توانند به متونی بسیار بیش از هر نوع پیکرهٔ گزینش‌شده‌ای دست یابند.

منابع متنی اینترنت همواره به‌روز هستند.

استفاده از منابع اینترنتی موجب صرفه‌جویی در زمان و هزینه‌های مادی می‌شود. زیرا نیازی به ایجاد پیکره‌ای مستقل نیست.

استفاده از اینترنت در مقایسه با پیکره این احتمال را افزایش می‌دهد که بتوان به واحدهای معنایی بیشتری برخورد کرد و اطلاعات زبانی بیشتری به‌دست آورد.

در دههٔ اخیر شاهد انقلابی بوده‌ایم که به‌نظر می‌رسد متأثر از پیشرفت ابزارهای قدرتمند اینترنتی، مانند موتورهای جست‌وجوی گوگل و پیمایشگرهای متمرکزی مانند بابوک بوده‌اند (**De Groc** 2011). این ادعا درست است. همان‌گونه که می‌دانیم، تعداد منابع متنی که هنگام جست‌وجوی فضای اینترنت در اختیار فرهنگ‌نویس قرار می‌گیرد بسیار زیاد و متنوع است و متناسب با تغییر جزئیات واژهٔ جست‌وجوشده، باز هم داده‌های بیشتری در اختیار تعریف‌نگار قرار می‌گیرد و این صرف‌نظر از شاهدهای بی‌اعتباری مانند شاهدهای مخدوش، تکراری، ماشینی و جز آن‌ها است. دسترسی به این حجم از منابع متنی به‌ویژه درمورد مدخل‌های معاصر، صورت‌های گفتاری، تفاوت‌های سبکی و کاربردشناختی و مانند آن‌ها بسیار قابل توجه و مفید است.

گستردگی شاهدها و تنوع بی‌نظیر متونی که با جست‌وجو در اینترنت، در مدت زمانی بسیار کوتاه در اختیار فرهنگ‌نویسان قرار می‌گیرند، این فرصت را فراهم می‌کند که در میان انبوهی از شاهدها به موارد زبانی دیگری، خارج از آنچه مستقیماً جست‌وجو شده‌است نیز برخورد کنند و به این ترتیب، به‌طور ضمنی فرهنگ‌نویس قادر می‌شود که دامنهٔ اطلاعات زبانی‌ای را که در فرهنگ به‌دست داده می‌شود گسترش دهد. برای نمونه، جست‌وجوی یک عنوان شغلی می‌تواند در زمانی بسیار کوتاه انبوهی از اطلاعات و اصطلاحات تخصصی

مربوط به آن پیشه یا حرفه را در اختیار فرهنگ‌نویس قرار دهد که به‌سختی می‌توان در متون مکتوب شاهدهایی برای آن یافت و جمع‌آوری میدانی آن‌ها نیز کاری زمان‌بر و پرهزینه است. نمونه‌ای دیگر مواردی است که جست‌وجو دربارهٔ یک واژه، فرهنگ‌نویس را با اشتقاق‌های جدیدی از آن مواجه می‌کند. نمونهٔ بسیار در دسترس آن، انبوه هم‌آیی‌ها با فعل‌های سبک و فعل‌های مرکّب برساخته‌ای است که به‌وفور در گفتار فارسی‌زبانان امروزه به‌کار می‌رود. از موارد دیگر می‌توان به صورت‌های املایی متفاوت با صورت صحیح معیار اشاره کرد که در پیکرهٔ متون شاهد ندارند و فرهنگ‌نویس با جست‌وجو در اینترنت به شاهدهایی از کاربرد طبیعی آن‌ها برمی‌خورد که گاهی حتی نیازمند برچسب‌های سبکی چون عوامانه یا برچسب‌های کاربردشناختی (مانند «در تداول ...») هستند.

مورد دیگر به‌روز بودن منابع اینترنتی است که امکان مطالعه و بررسی واژه‌هایی را فراهم می‌کند که به‌رغم آگاهی گویشوران از آن‌ها و به‌کارگیری آن‌ها، پیدا کردن شاهدهایی از کاربرد طبیعی‌شان در پیکره اغلب با دشواری بسیار همراه است؛ از آن جمله هستند اصطلاحات و عبارات کنایی که در تداول جوانان هر دوره به‌کار می‌رود و در گفتار و نوشتار غیررسمی در محیط‌های مجازی اینترنت یا شبکه‌های اجتماعی به‌وفور و با بسامد بالا به چشم می‌خورد.

نکتهٔ بعدی صرفه‌جویی در زمان و هزینه با استفاده از شاهدهای اینترنتی است. پوشیده نیست که تهیه و گردآوری پیکره بسیار زمان‌بر و نیازمند هزینه‌های مادی انبوه است. البته باید توجه داشت که استفاده از داده‌های اینترنتی برای تألیف فرهنگ، چیزی نیست که بتوان با اتکا به آن از پیکرهٔ مبتنی بر متن‌های منتشرشده چشم‌پوشی کرد. اما به‌عنوان بخشی از پیکره و در کنار شاهدهای برگرفته از منابع مکتوب، می‌توان از شاهدهای اینترنتی نیز بهره برد، بی‌آنکه زمان یا هزینه‌ای به پروژهٔ فرهنگ‌نویسی اضافه شود.

ترپ و فوئرتس ـ اولیورا (Tarp and Fuertes-Olivera 2016, p. 208) این موارد را جزو معایب استفادهٔ مستقیم از اینترنت به‌عنوان پیکره برمی‌شمرند:

کیفیت و اصالت منبع را در متن‌های برگرفته از اینترنت نمی‌توان کنترل کرد.
مؤلفان برخی متن‌ها اشخاص حقیقی نیستند.
گاهی نوع زبان مؤلفان برخی متن‌ها از تخصص، مهارت و دانش زبانی بی‌بهره است.
ممکن است متن‌ها ویرایش و بازبینی نشده باشند.
بسامدگیری از واژه‌های به‌کاررفته در متن‌های اینترنتی دشوار است.

در توضیح موارد بالا توجه به چند نکته ضروری است. نخست اینکه شاید استفاده از واژهٔ معایب چندان دقیق نباشد. موارد یادشده در بالا ناشی از ماهیت فضای اینترنت هستند و آنچه باید در نظر داشت این است که در برخورد با منابع اینترنتی لازم است حساسیت بیشتری به خرج داد و در استفاده از آن‌ها مدیریت و پایش داده‌ها را در نظر داشت. بنابراین، شاید صحیح‌تر باشد که موارد بالا را حساسیت‌های استفاده از منابع اینترنتی در نظر گرفت.

ازآنجاکه کاربران اینترنت، طیف بی‌شماری از افراد گوناگون هستند، بی‌گمان نمی‌توان از همهٔ متن‌های موجود در اینترنت شاهد آورد. برای نمونه، یکی از حساسیت‌هایی که در استفاده از منابع اینترنتی مطرح است، مسئلهٔ اصالت منبع و مؤلف آن است. نویسندهٔ بیشترِ شاهدهای برگرفته از صفحه‌های شخصی در اینترنت ناشناخته‌اند یا نام‌های غیرواقعی یا مستعار دارند و نمی‌توان به آن‌ها ارجاع داد. برخی شاهدها را نیز اشخاص حقیقی ننوشته‌اند یا بسیاری از شاهدها در صفحه‌ها و وبگاه‌های گوناگون عیناً تکرار شده‌اند. ازاین‌رو، بسامد واژه صرفاً عددی نیست که در صفحه‌های جست‌وجوگر نمایش داده می‌شود.

به‌علاوه ازآنجاکه فرهنگ‌نویسان درپیِ استفاده از شاهدهای گویا و صحیح‌اند و بیشترِ متن‌های اینترنتی ناویراسته‌اند، گاهی یافتن چنین شاهدهایی دشوار است.

افزون‌براین، شاهدهای اینترنتی تغییرپذیرند و گاه نمی‌توان به آن‌ها ارجاع داد. جست‌وجوگرها الگوریتم‌های پیچیده و عوامل متعددی را برای ترتیب نمایش نتایج به‌کار می‌برند، عواملی مانند بسامد بالای بازدید از یک وبگاه، منطقهٔ جغرافیایی که در آن جست‌وجو صورت گرفته، سابقهٔ جست‌وجوهای پیشین انجام‌شده از یک IP، صفحه‌های اندیس‌گذاری‌شده در آن موتور جست‌وجو، ممنوعیت‌های کشوری، رتبه‌بندی براساس کیفیت وبگاه‌ها، به‌روز بودن و امنیت وبگاه‌ها، شبکه‌های اجتماعی و محیط‌های کاربری، ترجمه‌ها، و کیفیت و تعداد پیوندهای وبگاه. برای نمونه، جست‌وجوگر گوگل بیش از ۲۰۰ عامل را برای رتبه‌بندی و ترتیب نمایش نتایج هر جست‌وجو بررسی و اعمال می‌کند. ازاین‌رو، گاه جست‌وجوی یک واژه در زمان‌ها و مکان‌های گوناگون نتیجهٔ متفاوتی دارد. با این توصیف، به‌راحتی نمی‌توان به شاهدهای اینترنتی ارجاع داد.

بی‌گمان شاهدهای برگرفته از پیکره قابل‌اعتماد و معتبرتر از شاهدهای برگرفته از اینترنت‌اند. در این میان این پرسش اساسی وجود دارد که با توجه به نوع فرهنگ و کاربران آن، به چه دامنه‌ای از واژگان نیاز است؟ آیا واژگان گفتاری، اصطلاحات یا تلفظ‌ها و صورت‌های املایی یا عوامانه، واژه‌های دارای برچسب‌های کاربردشناختی و جز آن‌ها در محدودهٔ واژگان فرهنگ هستند یا خیر. اگر پاسخ مثبت است، مانند آنچه در فرهنگ جامع

زبان فارسی شاهد آن هستیم، بهتر است با وجود حساسیت‌های یادشده، از شاهدهای اینترنتی بهره گرفته شود. زیرا منابع اینترنتی امکان دسترسی گسترده به شاهدهایی را فراهم می‌کنند که دستیابی به آن‌ها صرفاً با اتکاء به شاهدهای برگرفته از پیکره‌های رایج اگر ناشدنی نباشد، بسیار دشوار است.

۴ـ شاهدهای اینترنتی در فرهنگ جامع زبان فارسی

همان‌گونه که اشاره شد، فرهنگ جامع زبان فارسی نخستین فرهنگ فارسی است که در آن از شاهدهای اینترنتی در کنار شاهدهای برگرفته از پیکره بهره برده شده‌است. شیوهٔ استفادهٔ از شاهدهای اینترنتی فرهنگ جامع در ابتدا به‌صورت استفادهٔ مستقیم از داده‌ها و نتایج جست‌وجوگرهای اینترنتی بوده‌است. یعنی پیکره‌ای از داده‌های اینترنتی گردآوری نشده بوده‌است. در شیوه‌نامهٔ این فرهنگ (مقدمهٔ جلد ۱ و ۲) هرچند به استفاده از شاهدهای اینترنتی اشاره شده، اما شیوه‌نامه‌ای برای آن عرضه نشده‌است. به‌دلیل محدودیت‌های زمانی، اصولاً رویکرد فرهنگ جامع این بوده‌است که جست‌وجوی شاهد در اینترنت برای هر مدخل تنها زمانی ضرورت پیدا می‌کند که شاهدهای برگرفته از پیکره برای آن مدخل، به‌ویژه در دورهٔ معاصر، کامل نباشد. در جلد اول جست‌وجوی شاهدها تنها از طریق جست‌وجوگر گوگل و با تکیه بر تشخیص تعریف‌نگاران فرهنگ صورت می‌گرفته‌است.

همان‌طور که گودمان (Gudmann 2014, p. 31) به‌درستی گفته‌است هنوز هم انسان‌های واقعی در به‌صورت خلاقانه و بدون هیچ منبع ازپیش‌موجود و قطعی و هیچ پاسخ صحیح حک‌شده روی سنگ فرهنگ‌ها را تألیف می‌کنند. به عبارت دیگر، ابزارها و روش‌های نوین و بنیان‌ها و اصول مبتنی بر تجربه و منابع و داده‌های جدید هیچ‌کدام به‌تنهایی نمی‌توانند یک فرهنگ تألیف کنند. با وجود پیشرفت شگفت‌انگیز فناوری و به‌دنبال آن تغییر روش‌های فرهنگ‌نویسی، همچنان انسان مهم‌ترین عامل تألیف فرهنگ است، یعنی فرهنگ‌نویسان ماهر، دارای دانش زبانی، باتجربه و آموزش‌دیده.

البته باید به این نکته توجه داشت که در تألیف فرهنگ بین استفاده از توانش زبانی و به‌کارگیری دانش تخصصی تفاوت وجود دارد. کیلگاریف (Kilgarriff 1997, p. 111) معتقد است که برای تألیف فرهنگ و استخراج معنی از داده‌ها و براساس چگونگی رفتار واژه‌ها، تعریف‌نگار پیوسته بین دو منبع اولیهٔ تصمیم‌گیری و تشخیص در حرکت است؛ یعنی بین استفاده از توانش زبانی‌اش به‌عنوان یک گویشور زبان و استفاده از دانش تخصصی‌اش و رجوع به بافت به‌عنوان یک فرهنگ‌نویس.

ازاین‌رو، در تـألیف فرهنگ جامع نیز مانند دیگر فرهنگ‌ها در دیگر نقاط جهان، با تکیه بر دانش تعریف‌نگاران و دقت ویراسـتاران در بازبینی تعریف‌ها و شـاهدها، به جست‌وجوی شـاهد در اینترنت پرداخته شـده‌اسـت. نتیجهٔ این رویکرد چشمگیر بوده‌است و در جلد اول فرهنگ شـاهد آن هسـتیم و از آن جمله اسـت افزایش چشـمگیر تعداد برش‌های معنایی، صـورت‌های گفتاری و املایی متعدد، انواع اطلاعات کاربردشـناختی و سـبک‌شـناختی، ترکیبات، و همایندها.

پس از آن ویراسـتاران و مسـئولان گروه نظر به حسـاسـیت‌های اسـتفاده از شـاهدهای اینترنتی که پیش‌تر برشـمرده شـدند و از آن میان به‌ویژه با توجه به مسـئلهٔ کیفیت و اصـالت منابع، مشـکل مؤلفانِ شـاهدهای اینترنتی و سـطح زبان به‌کاررفته در این شـاهدها به لحاظ تخصـص و مهارت، دانش زبانی، و حتی گاهی صحت آن‌ها به‌لحاظ ملاک‌های دستوری و جز آن‌ها، اولویتی را به ملاک‌های جسـت‌وجوی شـاهد در اینترنت افزودند که براسـاس آن اولویـت بـه شـاهدهایی تعلق می‌گرفت که از منابع معتبری مانند مقالات موجود در وبگاه نورمگز اسـتخراج شـده باشند. به این ترتیب، تعریف‌نگاران از جلد ۲ فرهنگ موظف شدند اولویتشـان برای انتخاب شـاهد درمورد واژه‌های بی‌نشـان (نه صـورت‌های خاص مانند گفتاری، عوامانه و مانند آن یا صـورت‌هایی که در تداول گروه‌های خاص به‌کار می‌روند[1]) به شاهدهای به‌دست‌آمده از مقالات نورمگز تعلق گیرد.

هرچند رویکرد اخیر مسـئلهٔ حساسیت در کیفیت و اصالت منبع یا سطح زبان به‌کاررفته توسـط مؤلفان را برطرف می‌کند، اما گاه مسائلی نیز در این میان مغفول می‌ماند که در ادامه به آن‌ها خواهیم پرداخت. پیش از آن برای روشـن‌تر شـدن مطلب به مثال‌های زیر که جهت نمونه و به صورت تصادفی از فرهنگ جامع زبان فارسی انتخاب شده‌است توجه کنید:

ابتکار (در این جلد اولویت به شاهدهای برگرفته از نورمگز تعلق گرفته‌است)

در گوگل با صـورت فعلی ابتکار زدن روبه‌رو می‌شـویم و شـاهدها نشـان می‌دهد که در گفتار رسمی کاربرد ندارد و باید به آن برچسب «خودمانی» افزود. در نورمگز ابتکار بیش از سی‌هزار بار به‌کار رفته، ولی ابتکار زدن (نه «دست به ابتکار زدن») شاهدی ندارد.

---

۱. درمورد صـورت‌های خاص، مانند گفتاری و عوامانه، یا صـورت‌هایی که در تداول گروهی خاص به‌کار می‌رود، مانند اصطلاحات جدید جوانان و فضای مجازی که گاه حتی در پیکره هم شاهدی ندارند، تعریف‌نگار با تکیه بر شم زبانی و توانش خود به جست‌وجوی شاهد در اینترنت می‌پردازد و بی هیچ محدودیتی شاهد را انتخاب می‌کند.

آخرسر و آخرکاری

در این مدخل‌ها دو زیرمدخل آخرسرش و آخرکاری‌ها وجود دارد که شاهدهای آن‌ها برگرفته از اینترنت‌اند. ازآنجاکه این زیرمدخل‌ها در جلد اول قرار دارند، در هنگام تعریف‌نگاری‌شان از نورمگز استفاده نمی‌شده‌است، هرچند که در نورمگز نیز شاهد ندارد.

اتوبوسی

این مدخل در فرهنگ جامع زبان فارسی فوت شده‌است، اما برای نگاهی به تفاوت نتایج جست‌وجو، موردی خوب و قابل تأمل است. این واژه در گوگل شاهدهای بسیاری دارد: «قطار / نفربر اتوبوسی»، «ناوگان اتوبوسی»، «سفرهای / تصادفات اتوبوسی»، «دفاع اتوبوسی (در فوتبال)»، «کتابخانهٔ اتوبوسی»، «فیلم‌های اتوبوسی»، «جَک / تایر اتوبوسی»، «مدیریت / سیاست / آرای اتوبوسی»، «خانه‌ای با نقشهٔ اتوبوسی». ظاهراً دست‌کم با ۷ برش معنایی روبه‌رو هستیم.

بیشترِ شاهدهای این واژه در نورمگز «اتوبوس + ی (نکره)» هستند و چند شاهد نیز از «تبلیغاتِ اتوبوسی»، «مدیریت اتوبوسی» و «تایر اتوبوسی» یافت شد. بنابراین، ظاهراً دست‌کم با ۳ برش معنایی روبه‌رو هستیم.

نگاهی ساده به همین چند نمونه نشان می‌دهد که تفاوت تأمل‌برانگیزی بین این دو رویکرد جست‌وجو از نظر نتایج و شاهدهای به‌دست‌آمده وجود دارد.

باید توجه داشت که نورمگز پایگاهی برای ثبت و ارائهٔ مقالات حوزهٔ علوم انسانی و اسلامی است. در همین مرحله دو مسئلهٔ مهم وجود دارد؛ اول اینکه شاهدهای برگرفته از این وبگاه همگی از مقاله‌ها هستند. ازآنجاکه زبان مقاله‌نویسی و دایرهٔ واژگان به‌کاررفته در مقاله‌ها رسمی و علمی است، امکان مواجه شدن تعریف‌نگار با کاربردهای نشان‌دار محدود می‌شود. نیز ممکن است تعریف‌نگار در گام نخست متوجه کاربردهای نشان‌دار یا برش‌های معنایی خاص نشود، مانند آنچه در صورت فعلی ابتکار زدن در مدخل ابتکار به آن اشاره شد. این محدودیت می‌تواند بسیار تأثیرگذار باشد و احتمال از دست رفتن برخی اطلاعات از این دست را بالا ببرد.

دوم اینکه مقاله‌های این وبگاه نیز به لحاظ محتوا محدودیت دارند و تنها شامل مقالات حوزهٔ علوم انسانی و اسلامی هستند. همین مسئله نیز به نوعی دیگر محدودیت‌هایی را به‌طور ضمنی در این مقالات، به لحاظ سبک نگارش و دایرهٔ واژگان به‌کاررفته به همراه دارد.

باید توجه داشت که هرچند برخی شاهدهای اینترنتی به‌لحاظ اصالت و کیفیت منبع با شاهدهای پیکره یا شاهدهای برگرفته از مقاله‌های چاپ‌شده قابل مقایسه نیستند، اما شاهدهایی هستند که در بسیاری از موارد بازتاب زبان فارسی روزمره و زندهٔ امروزند و از این لحاظ بسیار ارزشمند و درخور توجه و بررسی هستند.

حساسیت در انتخاب شاهد نباید منجر به نوعی بنیادگرایی ادبی یا زبانی شود به گونه‌ای که شاهدهایی اولویت ورود به فرهنگ پیدا کنند که برگرفته از منابعی با شیوه‌ای خاص از نگارش یا با دایرهٔ واژگان و نوعی خاص از ادبیات باشند.

باید همواره در نظر داشت که فرهنگ در یک زبان محل قضاوت و گزینش شاهدها بر پایهٔ معیارهایی مانند کیفیت ادبی متن یا سطح دانش مؤلف آن نیست. در فرهنگ‌ها، زبان با تمام کارکردهای آن و در تمام سطوح سبکی از اعتبار واحدی برخوردار است، زیرا همهٔ این موارد کاربردهای طبیعی و اصیل زبان هستند و امتیازدهی به متون از لحاظ شیوایی و زیبایی یا ارزش ادبی یا دسته‌بندی‌های سبکی، در انتخاب شاهدهای فرهنگ نباید دخیل باشند.

به علاوه این نکته را هم نباید از نظر دور داشت که چنین قضاوت‌هایی باعث می‌شود بخشی از زبان نادیده گرفته شود و شاهدهای فرهنگ، تصویری از زبان ارائه دهند که کامل نیست. اگر اولویت انتخاب شاهدها صرفاً به مقالات علمی داده شود، این خطر وجود دارد که برای نمونه کاربرد واژگانی که به لحاظ سبکی بیشتر در متون یا مقالات علمی به‌کار می‌روند، در فرهنگ برجسته‌تر و پربسامدتر از واقعیت شود.

همچنین ویراستاران شاهدهایی را که تعریف‌نگاران انتخاب می‌کنند، در مرحله‌ای دیگر بازبینی می‌کنند و در برخی موارد از میان این شاهدها نیز، تنها تعدادی در اولویت برای انتخاب نهایی قرار می‌گیرند که از مؤلفانی شناخته‌شده‌تر و یا از نشریاتی معتبرتر استخراج شده باشند. مجموعهٔ این محدودیت‌ها می‌تواند، نه در همه موارد، اما در برخی مدخل‌ها باعث از دست رفتن برخی اطلاعات و یا در نگاهی کلان به مجموعه‌ای از شاهدهای فرهنگ و بسامد واژگان این شاهدها، به نتایجی متفاوت با واقعیت منجر شود.

بنابراین، به نظر می‌رسد هرچند که رویکرد جدید فرهنگ جامع به رفع حساسیت درمورد کیفیت و اصالت منابع شاهدها یا سطح زبان و واژگان به‌کاررفته توسط مؤلفان کمک می‌کند، اما نادیده گرفتن برخی نکته‌ها آسیب‌هایی برای فرهنگ به دنبال داشته‌است.

نتیجه‌گیری

تعریف‌نگاران فرهنگ‌ها، در کنار پیکره‌ها، نیازمند اســتفاده از داده‌های اینترنتی نیز هســتند. البته استفاده از شاهدهای اینترنتی در کنار مزایای چشمگیر آن، حساسیت‌ها و مشکلاتی نیز دارد. با وجود تمام این مشـکلات، چشم‌پوشی از شاهدهای اینترنتی درست و علمی نیست و ازآنجاکه تشـخیص معنی واحدهای واژگانی، با توجه به حشو موجود در بافت‌های طبیعی زبان و مســائل بســط اســتعاری و مجازی معنا، دشــوار اســت، بخش خلاق بهره‌گیری از شـاهدهای اینترنتی یعنی توصیف و تحلیل معنی و طبقه‌بندی بافت‌ها و شاهدهای مورد نیاز برای فرهنگ‌نویسی، همچنان به عهدهٔ تعریف‌نگاران و ویراستاران فرهنگ است.

منابع

خطیبی، ابوالفضـل (۱۳۸۶)، «فرهنگ جامع زبان فارسی، پیکره در فرهنگ نویسـی فارسی و پیکره زبانی رایانه‌ای»، فرهنگ‌نویسی، شمارهٔ ۱، صفحه‌های ۴ـ۶۷.

فرهنگ جامع زبان فارسی (۱۳۹۲و۱۳۹۵)، زیر نظر علی‌اشرف صادقی، فرهنگستان زبان و ادب فارسی، تهران.

De Groc, C. (2011, August). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 497-498). IEEE.

Fuertes-Olivera, P. A. (2012), "Lexicography and the Internet as a (Re-) source", *Lexicographica*, 28(1), 49-70.

Gudmann, H. R. (2014), Betydningshuller i Spanske Ordbøger. En Undersøgelse af Betydningsenheder i Spanske Monolingvale Almene Receptionsordbøger, M.A. Thesis. Aarhus: Aarhus University, Department of Business Communication.

Kilgarriff, A. (1997), "I don't believe in word senses", *Computers and the Humanities*, 31(2), 91-113.

Kilgarriff, A. and G. **Grefenstette** (2003), "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics* 29(3): 333-347.

Tarp, S. T., and **Fuertes-Olivera**, P. A. (2016), "Advantages and disadvantages in the use of internet as a corpus: The case of the online dictionaries of Spanish Valladolid-UVa", *Lexikos*, 26(1), 273-295.

Zgusta, L. (1989), "Probable future developments in lexicography", *Hausmann, FJ et al.(Eds.)*, 1991, 3157-3167.