

## بررسی مزایا و معایب استفاده از منابع اینترنتی به عنوان بخشی از پیکره فرهنگ نویسی

سارا شریف پور (عضو هیئت علمی فرهنگستان زبان و ادب فارسی)

چکیده: در این مقاله استفاده از شاهد‌های اینترنتی در فرایند فرهنگ نویسی بررسی شده است و به این منظور شاهد‌هایی از فرهنگ جامع زبان فارسی ارائه شده که نشان می‌دهد مؤلفان این فرهنگ در کنار بهره‌مندی از پیکره زبانی عظیم خود، به دلیل نزدیکی شاهد‌های اینترنتی به کاربرد روزمره و طبیعی زبان، نیازمند استفاده از این شاهد‌ها نیز هستند. با توجه به این نیاز، در مقاله پیش رو با استفاده از تحلیل ترپ و فوئرتس - اولیورا (۲۰۱۶)، مزایا و حساسیت‌های استفاده از شاهد‌های اینترنتی معرفی شده و به بررسی نقش فعال مؤلفان فرهنگ در استفاده بهینه و مؤثر از داده‌های اینترنتی پرداخته شده است. در ادامه، براساس این تحلیل، رویکردهای اتخاذ شده در فرهنگ جامع زبان فارسی برای استفاده از شاهد‌های اینترنتی بررسی شده و با به دست دادن نمونه‌هایی از مدخل‌های این فرهنگ، آن‌ها آسیب‌شناسی و تحلیل شده است. کلیدواژه‌ها: پیکره فرهنگ نویسی، پیکره اینترنتی، شاهد‌های زبانی اینترنتی، استخراج شاهد‌های اینترنتی

### ۱- مقدمه

پایه و اساس تألیف هر فرهنگی را پیکره آن تشکیل می‌دهد. پیکره مجموعه‌ای از داده‌های نوشتاری یا گفتاری است که فرهنگ‌نویسان از آن‌ها برای استخراج اطلاعات گوناگون و متعددی درباره 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 محدوده کاربری تعریف شده و کاربران مفروض آن دارد.

عموماً پیکره‌های زبانی شامل منابع زیر هستند: منابع نوشتاری که می‌تواند از کهن‌ترین نوشته‌ها تا زمان تألیف فرهنگ، شامل کتاب‌ها و مقاله‌هایی در موضوعات گوناگون، نامه‌ها، سفرنامه‌ها و غیره را دربر بگیرد. منابع گفتاری که حاوی گفت‌وگوهای روزمره اهل زبان، به‌ویژه برای ثبت تلفظ صحیح یا امروزی واژه‌ها، است و منابع حاصل از پژوهش‌های میدانی که به‌ویژه در مورد اصطلاحات مربوط به حرفه‌ها و پیشه‌ها کاربرد دارد (خطیبی ۱۳۸۶، ص ۱۴).

خطیبی (۱۳۸۶) در مقاله خود، به معرفی و بررسی نقش پیکره و کهن‌ترین فرهنگ‌های فارسی از حدود قرن ۵ هجری تا دوران معاصر پرداخته است. آنچه از این پژوهش دریافت می‌شود این است که فرهنگ‌ها عموماً متشکل از منابع نوشتاری، که سهم قابل توجهی از آن‌ها به پیکره متون منظوم و منثور زبان فارسی اختصاص داشته است و نیز در برخی موارد منابع حاصل از پژوهش‌های میدانی بوده است. این مسئله شامل فرهنگ‌های تألیف‌شده در دوره معاصر هم می‌شود.

در میان فرهنگ‌های معاصر فارسی، فرهنگ جامع زبان فارسی توانسته است به اقتضای گسترش فضای مجازی و دنیای اینترنت در سراسر جهان و همسو با سیاست سایر فرهنگ‌ها در کشورهای دیگر، از این فضا به‌عنوان منبع زبانی جدید و غنی بهره‌برد که امری بی‌سابقه در تاریخچه فرهنگ‌نویسی در زبان فارسی است.

## ۲- منابع اینترنتی به‌عنوان بخشی از پیکره

حضور رایانه در مراحل گوناگون فرهنگ‌نویسی را شاید بتوان همانند انقلابی در این حوزه در نظر گرفت. بارزترین نمود این حضور، جنبه فنی و وجه کاربرد ابزاری رایانه در فرهنگ‌نویسی است که امتیاز اصلی آن، توانایی در سامان‌دهی سریع داده‌ها، حفظ انسجام و ترتیب آن‌ها و بازسازی آن‌ها در قالب زیرمجموعه‌ها (پیوندها و ارجاعات و غیره) با استفاده از نرم‌افزارهای فرهنگ‌نویسی است. همچنین گردآوری همایندهای موجود در بافت‌ها 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 برجسته رایانه‌ها در حوزه انتشار فرهنگ‌هاست که موجب صرفه‌جویی در منابع مادی و افزایش چشمگیر سرعت و زمان مورد نیاز برای چاپ یا ویرایش فرهنگ‌ها می‌شود (Zgusta 1989).

اما ورود رایانه به حوزه معنی به اندازه جنبه‌های فنی و ابزاری سریع و قدرتمند نبوده است. تشخیص معنای ناشناخته واحدهای واژگانی بر پایه انبوه داده‌های زبانی، با توجه به حشو موجود در بافت‌های طبیعی زبان و مسائل بسط استعاری و مجازی معنی امری دشوار است که هنوز رایانه قادر به انجام دقیق آن نیست. به عبارت دیگر، توصیف و تحلیل معنی و طبقه‌بندی بافت‌ها و شاهد‌های مورد نیاز در فرهنگ‌نویسی یا به تعبیر دقیق‌تر، بخش خلاق بهره‌گیری از داده‌ها، عمدتاً همچنان به‌عهده تعریف‌نگاران و ویراستاران فرهنگ است. با وجود این، نباید از نقش تأثیرگذار و البته کمی رایانه‌ها در حوزه معنی و مشخصاً در گردآوری پیکره‌ها غافل شد که به‌صورت غیرمستقیم و با فراهم کردن امکان دسترسی به انبوهی از داده‌ها، کمک چشمگیری به مؤلفان فرهنگ در تشخیص و تفکیک معنی می‌کند.

فوئرتس - الیورا (Fuertes-Olivera 2012, p. 51) معتقد است که پیکره فرهنگ‌نویسی را (یعنی پیکره‌ای که بتوان از آن در تألیف فرهنگ بهره برد) می‌توان به این صورت تعریف کرد: هر مجموعه‌ای از متون (نوشتاری یا گفتاری) که فرهنگ‌نویسان از آن برای تکمیل ساختار فرهنگ و تألیف آن بهره می‌برند. در نتیجه، متون موجود در فضای اینترنت را هم به همین تعبیر می‌توان مجموعه‌ای از متون در نظر گرفت که بر اساس این تعریف می‌تواند به‌عنوان پیکره فرهنگ‌نویسی به‌کار گرفته شود. این همان نظری است که کیلگاریف و گرنستت (Kilgarriff and Grefenstette 2003, p. 334) نیز آن را تأیید کرده‌اند و به این پرسش که آیا اینترنت را می‌توان یک پیکره در نظر گرفت، به‌روشنی پاسخ مثبت داده‌اند.

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

در کشور ما، برخلاف دیگر کشورها، استفاده از اینترنت به‌عنوان پیکره فرهنگ‌نویسی یا بخشی از آن هنوز چندان به‌کار گرفته نشده یا حتی تأیید نشده است. از این رو، در این رابطه می‌توان فرهنگ جامع زبان فارسی را نمونه‌ای پیشرو در کشور ایران در نظر گرفت که علاوه بر پیکره زبانی عظیم خود که مبتنی است 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 برده است. این ویژگی وجه ممیز بسیار ارزشمندی است که این فرهنگ را از سایر فرهنگ‌های معاصر متمایز می‌کند. از میان فرهنگ‌های فارسی معاصر، شاید تنها فرهنگ بزرگ سخن این امکان را داشته است که از منابع اینترنتی هم به‌عنوان بخشی از پیکره بهره برد که البته این اتفاق نیفتاده و مقایسه آن با فرهنگ جامع زبان فارسی به‌روشنی گویای تأثیر

آشکار منابع متنی اینترنت در غنای بیشتر امکانات فرهنگ‌نویسی به‌ویژه درج معنی‌های بیشتر، مواجهه با صورت‌های املائی، گفتاری، عوامانه و جز آن‌ها و نیز دسترسی به انواع تفاوت‌های سبکی است. مدخل‌های زیر به‌طور تصافی از جلد اول فرهنگ جامع زبان فارسی انتخاب شده‌اند. اطلاعاتی که برای هر مدخل در زیر برشمرده شده‌است، همگی مواردی هستند که تنها با تکیه بر شاهد‌های اینترنتی به‌دست آمده و در فرهنگ درج شده‌اند. مشابه این اطلاعات در هیچ‌یک از فرهنگ‌های فارسی موجود، حتی فرهنگ بزرگ سخن، که معنی‌ها و شاهد‌های معاصر آن پررنگ است، وجود ندارد:

### فرهنگ جامع زبان فارسی - جلد ۱

<p><b>آزانی</b> یک. (صفت) مربوط به آژانس. ... دو. (اسم) کسی که در آژانس کار می‌کند.</p> <p><b>آس</b><sup>۴</sup> دو. (صفت) (مجاز) بسیار عالی و بی‌نظیر.</p> <p><b>آقای گل</b>: (ورزش) (معمولاً در فوتبال) بازیکنی که در مسابقه‌های تیم خود با تیم‌های دیگر، بیشترین تعداد گل را به بازیکنان تیم‌های رقیب زده باشد.</p> <p><b>آک</b><sup>۲</sup> صورت کوتاه آکبند (ص ۸۳۰).</p> <p><b>آتا</b> یک. (صفت) (کسی یا چیزی) که از نظر برنامه‌ریزی و زمان‌بندی، منظم و سر وقت باشد.</p>	<p><b>آره و آجرپاره</b> (شبه‌جمله) (خودمانی) (برای تحقیر و توهین) (مجاز) در پاسخ به جواب «آره» کسی و برای اعتراض به آن همراه با عصبانیت و سرزنش به‌کار می‌رود.</p> <p><b>آره</b><sup>۳</sup> معنی ۹: با لحن پرسش‌شی، برای نشان دادن تعجب در برابر گفته دیگری به‌کار می‌رود.</p> <p><b>آزمایش</b> آزمایش دادن: به آزمایشگاه مراجعه کردن و خود را تحت آزمایش پزشکی قرار دادن.</p> <p><b>آزمایشی</b> دو. (قید) (درمورد انجام گرفتن امری) به‌طور موقت تا کیفیت و کارایی آن برای هدفی مشخص، مورد ارزیابی قرار گیرد.</p>
--	--

چنان‌که دیده می‌شود، دسترسی به داده‌های اینترنتی باعث شده‌است که در هر مدخل اطلاعات و جزئیات بیشتری ارائه شود که به‌ویژه از نظر تعداد برش‌های معنایی قابل تأمل است. این نکته را نیز باید در نظر داشت که نمونه‌های یادشده «تنها شامل مواردی بودند که درج یک برش معنایی جدید یا ایجاد یک زیرمدخل و اطلاعات دیگری مانند آن، صرفاً مبتنی بر منابع اینترنتی بوده‌اند و از آوردن نمونه برای مواردی که در کنار شاهد‌های پیکره از شاهد‌های اینترنتی هم استفاده شده، صرف‌نظر شده‌است». ناگفته پیداست که در چنین مواردی نیز بهره‌گیری از منابع متنی اینترنت هرچند به درج اطلاعاتی جدید منجر نشده‌است، اما بی‌تردید بر غنای شاهد‌ها و اطلاعات ارائه‌شده در فرهنگ افزوده‌است.

اما آیا استفاده از منابع متنی اینترنت، در هر شکل و با هر کیفیتی مطلوب است؟ فرهنگ‌نویسان در این باره چه قاعده‌هایی دارند؟ و اینکه مؤلفان فرهنگ جامع در این باره چگونه عمل کرده‌اند؟

### ۳- 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 پیکره فرهنگ بررسی می‌شود تا امکان بررسی عملکرد فرهنگ جامع زبان فارسی نیز فراهم شود.

ترپ و فوئرتس - 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 مقاله 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 پیکره‌گزینش‌شده‌ای 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 دهه اخیر شاهد انقلابی بوده‌ایم که به نظر 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 واژه 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 در مورد مدخل‌های معاصر، صورت‌های گفتاری، تفاوت‌های سبکی و کاربردشناختی و مانند آن‌ها بسیار قابل توجه و مفید است.

گسترده‌گی 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 دامنه 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 مرگب بر ساخته‌ای است که به‌وفور در گفتار فارسی‌زبانان امروزه به‌کار می‌رود. از موارد دیگر می‌توان به صورت‌های املائی 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 به‌پروردهٔ فرهنگ‌نویسی اضافه شود.

ترپ و فوئرتس - 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 واژه 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. از آنجاکه کاربران اینترنت، طیف بی‌شماری از افراد گوناگون هستند، بی‌گمان نمی‌توان از همه متن‌های موجود در اینترنت شاهد آورد. برای نمونه، یکی از حساسیت‌هایی که در استفاده از منابع اینترنتی مطرح است، مسئله اصالت منبع و مؤلف آن است. نویسنده بیشتر 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 نوشته‌اند یا بسیاری از شاهد‌ها در صفحه‌ها و وبگاه‌های گوناگون عیناً تکرار شده‌اند. از این رو، بسامد واژه صرفاً عددی نیست که در صفحه‌های جست‌وجوگر نمایش داده می‌شود.

به‌علاوه از آنجاکه فرهنگ‌نویسان در پی استفاده از شاهد‌های گویا و صحیح‌اند و بیشتر 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، منطقه جغرافیایی که در آن جست‌وجو صورت گرفته، سابقه 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. از این رو، گاه جست‌وجوی یک واژه در زمان‌ها و مکان‌های گوناگون نتیجه 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 محدوده 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.

#### ۴- 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 شده است. شیوه استفاده از شاهد‌های اینترنتی فرهنگ جامع در ابتدا به صورت استفاده مستقیم از داده‌ها و نتایج جست‌وجوگرهای اینترنتی بوده است. یعنی پیکره‌ای از داده‌های اینترنتی گردآوری نشده بوده است. در شیوه‌نامه این فرهنگ (مقدمه جلد ۱ و ۲) هرچند به استفاده از شاهد‌های اینترنتی اشاره شده، اما شیوه‌نامه‌ای برای آن عرضه نشده است. به دلیل محدودیت‌های زمانی، اصولاً رویکرد فرهنگ جامع این بوده است که جست‌وجوی شاهد در اینترنت برای هر مدخل تنها زمانی ضرورت پیدا می‌کند که شاهد‌های برگرفته از پیکره برای آن مدخل، به ویژه در دوره معاصر، کامل نباشد. در جلد اول جست‌وجوی شاهد‌ها تنها از طریق جست‌وجوگر گوگل و با تکیه بر تشخیص تعریف‌نگاران فرهنگ صورت می‌گرفته است.

همان‌طور که گودمان (Gudmann 2014, p. 31) به درستی گفته است هنوز هم انسان‌های واقعی در به صورت خلاقانه و بدون هیچ منبع از پیش موجود و قطعی و هیچ پاسخ صحیح حک شده 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 به تنهایی نمی‌توانند یک فرهنگ تألیف کنند. با وجود پیشرفت شگفت‌انگیز فناوری و به دنبال 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 اولیه 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.



از این رو، در تألیف فرهنگ جامع نیز مانند دیگر فرهنگ‌ها در دیگر نقاط جهان، با تکیه بر دانش تعریف‌نگاران و دقت ویراستاران در بازبینی تعریف‌ها و شاهدها، به جست‌وجوی شاهد در اینترنت پرداخته شده است. نتیجه این رویکرد چشمگیر بوده است و در جلد اول فرهنگ شاهد آن هستیم و از آن جمله است افزایش چشمگیر تعداد برش‌های معنایی، صورت‌های گفتاری و املائی متعدد، انواع اطلاعات کاربردی شناختی و سبک‌شناختی، ترکیبات، و هماینها.

پس از آن ویراستاران و مسئولان گروه نظر به حساسیت‌های استفاده از شاهدهای اینترنتی که پیش‌تر برش‌مرده شدند و از آن میان به‌ویژه با توجه به مسئله کیفیت و اصالت منابع، مشکل مؤلفان شاهدهای اینترنتی و سطح زبان به‌کاررفته در این شاهدها به لحاظ تخصص و مهارت، دانش زبانی، و حتی گاهی صحت آن‌ها به لحاظ ملاک‌های دستوری و جز آن‌ها، اولویتی را به ملاک‌های جست‌وجوی شاهد در اینترنت افزودند که براساس آن اولویت به شاهدهایی تعلق می‌گرفت که از منابع معتبری مانند مقالات موجود در وبگاه نورمگز استخراج شده باشند. به این ترتیب، تعریف‌نگاران از جلد ۲ فرهنگ موظف شدند اولویتشان برای انتخاب شاهد در مورد واژه‌های بی‌نشان (نه صورت‌های خاص مانند گفتاری، عوامانه و مانند آن یا صورت‌هایی که در تداول گروه‌های خاص به کار می‌روند<sup>۱</sup>) به شاهدهای به‌دست‌آمده از مقالات نورمگز تعلق گیرد.

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

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

در گوگل با صورت فعلی ابتکار زدن روبه‌رو می‌شویم و شاهدها نشان می‌دهد که در گفتار رسمی کاربرد ندارد و باید به آن برچسب «خودمانی» افزود. در نورمگز ابتکار بیش از سی‌هزار بار به کار رفته، ولی ابتکار زدن (نه «دست به ابتکار زدن») شاهی ندارد.

۱. در مورد صورت‌های خاص، مانند گفتاری و عوامانه، یا صورت‌هایی که در تداول گروهی خاص به کار می‌رود، مانند اصطلاحات جدید جوانان و فضای مجازی که گاه حتی در پیکره هم شاهی ندارند، تعریف‌نگار با تکیه بر شم زبانی و توانش خود به جست‌وجوی شاهد در اینترنت می‌پردازد و بی‌هیچ محدودیتی شاهد را انتخاب می‌کند.

### آخرسر و آخرکاری

در این مدخل‌ها دو زیرمدخل آخرسرش و آخرکاری‌ها وجود دارد که شاهد‌های آن‌ها برگرفته از اینترنت‌اند. از آنجاکه 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)»، «کتابخانه اتوبوسی»، «فیلم‌های اتوبوسی»، «جک / تایر اتوبوسی»، «مدیریت / سیاست / آرای اتوبوسی»، «خانه‌ای با نقشه اتوبوسی». ظاهراً دست‌کم با ۷ برش معنایی روبه‌رو هستیم.

بیشتر شاهد‌های این واژه در نورمگز «اتوبوس + ی (نکره)» هستند و چند شاهد نیز از «تبلیغات 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 ارائه مقالات حوزه علوم انسانی و اسلامی است. در همین مرحله دو مسئله مهم وجود دارد؛ اول اینکه شاهد‌های برگرفته از این وبگاه همگی از مقاله‌ها هستند. از آنجاکه زبان مقاله‌نویسی و دایرة‌واژگان 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 حوزه علوم انسانی و اسلامی هستند. همین مسئله نیز به نوعی دیگر محدودیت‌هایی را به‌طور ضمنی در این مقالات، به لحاظ سبک نگارش و دایرة‌واژگان به‌کاررفته به همراه دارد.

باید توجه داشت که هرچند برخی شاهد‌های اینترنتی به لحاظ 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 زنده‌امروزد 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 دایره‌واژگان و نوعی خاص از ادبیات باشند.

باید همواره در نظر داشت که فرهنگ در یک زبان محل قضاوت و گزینش شاهد‌ها بر پایه 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 همه 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 به کار 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. مجموعه 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 در مورد 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 استفاده از داده‌های اینترنتی نیز هستند. البته استفاده از شاهد‌های اینترنتی در کنار مزایای چشمگیر آن، حساسیت‌ها و مشکلاتی نیز دارد. با وجود تمام این مشکلات، چشم‌پوشی از شاهد‌های اینترنتی درست و علمی نیست و از آنجاکه تشخیص معنی واحدهای واژگانی، با توجه به حشو موجود در بافت‌های طبیعی زبان و مسائل بسط استعاری و مجازی معنا، دشوار است، بخش خلاق بهره‌گیری از شاهد‌های اینترنتی یعنی توصیف و تحلیل معنی و طبقه‌بندی بافت‌ها و شاهد‌های مورد نیاز برای فرهنگ‌نویسی، همچنان به عهده‌ی تعریف‌نگاران و ویراستاران فرهنگ است.

### منابع

- خطیبی، ابوالفضل (۱۳۸۶)، «فرهنگ جامع زبان فارسی، پیکره در فرهنگ‌نویسی فارسی و پیکره‌زبانی رایانه‌ای»، فرهنگ‌نویسی، شماره ۱، صفحه‌های ۴-۶۷.
- فرهنگ جامع زبان فارسی (۱۳۹۲ و ۱۳۹۵)، 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.