CLIP — portal internetowy łączący projekty, narzędzia, zespoły związane z komputerowym przetwarzaniem języka polskiego

CLIP (od ang. Computational Linguistics in Poland) to dostępna pod adresem http://clip.ipipan.waw.pl anglojęzyczna witryna internetowa, której celem jest informowanie międzynarodowej społeczności naukowej o aktywności środowiska lingwistyczno-informatycznego w Polsce. Witryna zawiera możliwie aktualne i kompletne informacje o zasobach, narzędziach, projektach i zespołach związanych z komputerowym przetwarzaniem polszczyzny w kraju i za granicą. CLIP kontynuuje tradycje strony o podobnym charakterze prowadzonej w latach 2000–2004 w Instytucie Podstaw Informatyki PAN, wprowadzając rozproszony model zarządzania informacją, znany np. z Wikipedii, polegający na udzieleniu prawa do modyfikacji stron wszystkim chętnym przedstawicielom społeczności naukowej (zwykle kierownikom grup badawczych i wskazanym przez nich osobom, np. kierownikom projektów).

Najważniejsza część strony CLIP zawiera obecnie wykaz przeszło 160 zasobów i narzędzi lingwistycznych powstałych w Polsce lub dotyczących języka polskiego. Wszystkie, co jest ważnym założeniem projektowym listy, można bezpośrednio pobrać z Internetu bez konieczności dodatkowej interakcji z ich właścicielami. Dostępne zasoby i narzędzia zostały pogrupowane tematycznie:

1) zasoby i narzędzia korpusowe — korpusy języka pisanego i narzędzia korpusowe, korpusy języka mówionego, korpusy równoległe i pamięci tłumaczeniowe,

2) zasoby i narzędzia słownikowe — słowniki komputerowe, narzędzia i zasoby morfologiczne, tagery (narzędzia do ujednoznaczniania morfoskładniowego),

3) inne — parsery, gramatyki, banki rozbiorów gramatycznych, narzędzia do analizy i syntezy mowy, systemy tłumaczeniowe, pozostałe.

Krótka informacja o zasobie (nazwa, kilkuwyrazowy opis, lista nazwisk twórców) opatrzona jest odnośnikiem do jego strony internetowej — zewnętrznej, stworzonej zwykle przez macierzystą jednostkę lub zintegrowanej ze stronami CLIP. W ten ostatni sposób właściciele praw autorskich mogą powierzyć opiekę nad stroną zasobu administratorom witryny CLIP, by np. uniknąć obsługi informatycznej związanej z utrzymywaniem własnej strony. W ten sposób częścią witryny stała się m.in. witryna Korpusu Frekwencyjnego Polszczyzny lat 60. XX w. (http://clip.ipipan.waw.pl/ PL196x), przeniesiona z serwera Wydziału Matematyki, Informatyki i Mechaniki UW.

Osobna podstrona CLIP zawiera odnośniki do stron internetowych 17 polskich zespołów naukowych zajmujących się inżynierią lingwistyczną (obecnie z Gdańska, Gliwic, Krakowa, Łodzi, Poznania, Warszawy i Wrocławia), zaś podstrona projektów uporządkowane chronologicznie informacje o zewnętrznie finansowanych projektach z tej dziedziny — od 1993 r. aż po obecnie trwające (do 2018 r.). Opisy projektów, z założenia znormalizowane, obejmują polską i angielską nazwę projektu, jego typ, źródła finansowania, okres realizacji oraz krótkie streszczenie.

W ciągu ostatniego roku działania witryny odwiedziło ją ponad 2500 osób, głównie z Polski (ok. 68% odwiedzin), ale także z Niemiec, Francji, Stanów Zjednoczonych i Wielkiej Brytanii. Witryna była wykorzystywana m.in. przez Komisję Europejską, zagraniczne firmy zajmujące się tłumaczeniem maszynowym i zagranicznych slawistów.

Portal CLIP jest jednym z wyników projektu CESAR, w którego ramach stworzono, rozszerzono i ustandaryzowano wiele narzędzi dla polszczyzny oraz opracowano tzw. raport językowy, prezentujący pozycję języka polskiego wśród innych języków europejskich pod kątem dostępności i zaawansowania narzędzi informatycznych do przetwarzania lingwistycznego (por. http://www.meta-net.eu/whitepapers/volumes/polish). W związku z zakończeniem projektu w pierwszej połowie 2013 r. i wynikającym z tym brakiem osobnego finansowania na aktywny rozwój witryny serdecznie zachęcamy polskie środowisko naukowe zajmujące się komputerowym przetwarzaniem języka do włączenia się w jej tworzenie. W celu uzyskania uprawnień do edycji prosimy o kontakt z autorami niniejszego tekstu.

Maciej Ogrodniczuk
Adam Przepiórkowski
Warszawa, Instytut Podstaw Informatyki PAN