DGT-UD: a Parallel 23-language Parsebank

Nikola Ljubešić, Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute
SI-1000 Ljubljana, Slovenia
{nikola.ljubesic,tomaz.erjavec}@ijs.si

The JRC DGT corpus

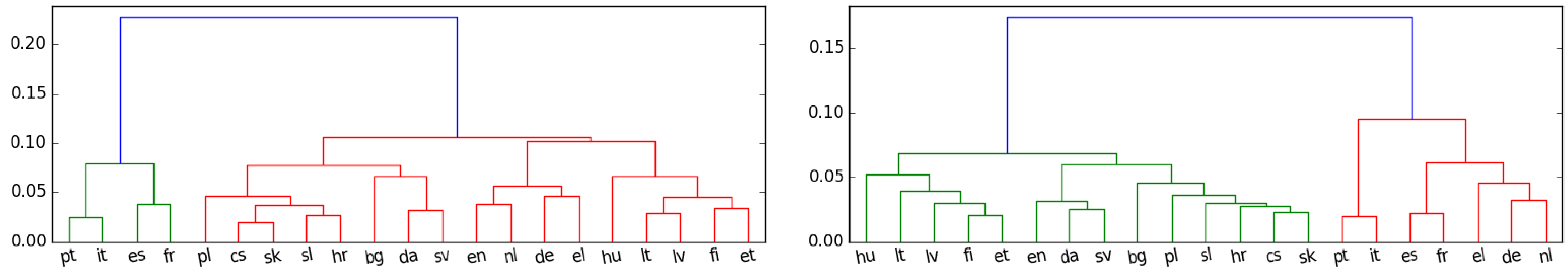
- JRC DGT corpus contains the translation memory of Acquis Communautaire (European Union law)
- Made publicly available in 2007 by the European Commission's Directorate-General for Translation (DGT) and the Joint Research Centre (JRC) of the European Commission
- Contains texts in 24 languages aligned on the sentence level and is distributed as a set of TMX files identified by the EUR-Lex number of the underlying documents.
- Proof-read texts, carefully checked translations, no duplicate sentences, is highly multilingual, large, available under a very permissive licence, and regularly updated.
- To create the DGT-UD corpus we took all the releases of JRC DGT, up to 2017.

UDPipe

- UDPipe 1.2 – tool developed by UFAL to be trained on the Universal Dependencies (UD) data, used as baseline in shared tasks run on UD data
- Performs tokenization, sentence segmentation, morphosyntactic tagging, lemmatization and parsing, all learned from UD data
- Covers 68 models of 50 languages (UD v2.0), 23 out of 24 DGT languages, Maltese the only unsupported language
- LAS scores: Bulgarian (84.8), Croatian (77.9), Czech (83.2), Danish (74.7), Dutch (69.6), English (77.2), Estonian (65.6), Finnish (76.9), French (80.7), German (68.6), Greek (80.4), Hungarian (68.1), Irish (62.4), Italian (86.1), Latvian (62.8), Lithuanian (32.4), Polish (80.7), Portuguese (82.6), Romanian (80.2), Slovak (75.2), Slovenian (81.6), Spanish (81.4), Swedish (77.8)

Comparative linguistic analysis

- **Question**: Is the automatically annotated DGT-UD better suited for comparative analyses than the manually annotated UD data?
- **Method**: Each CLARIN language represented as probability distribution of UD label trigrams, hierarchical clustering of languages.
- **Figure**: Dendrograms representing clustering results on DGT-UD (left) and UD data (right)



Structure

The corpus consists of 1 alignment file, 23 vertical files and registry files (as used by the Manatee corpus manager back-end)
Structural attributes of vertical files:

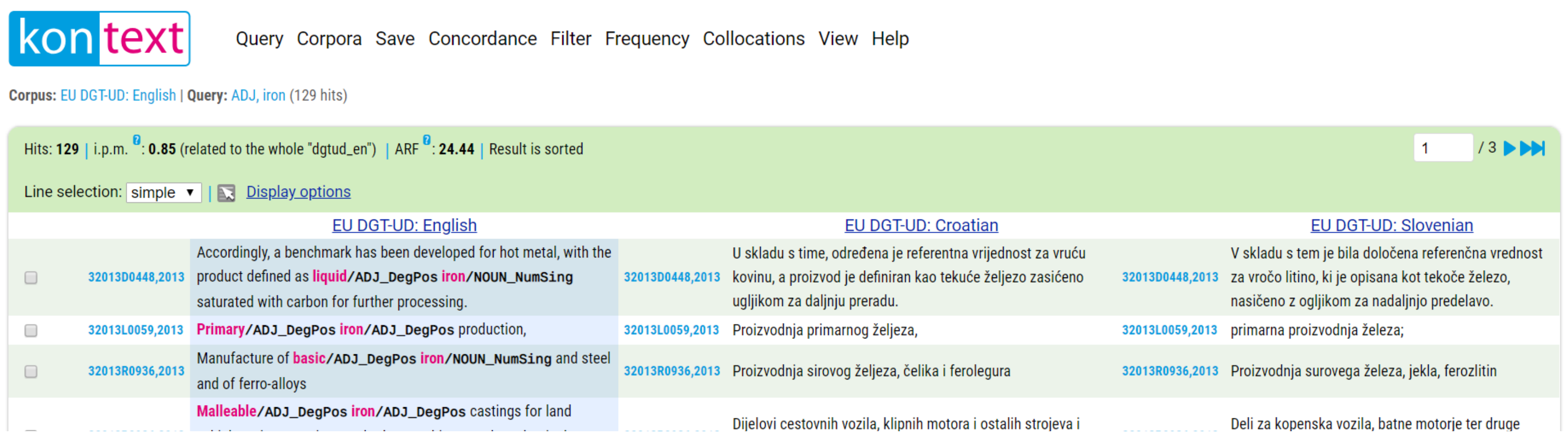
- *Text* with id, year of publication & release e.g. `<text id="213A10" year="-2003" release="DGT-TM-2007">`
- *Anonymous block*, the unit of alignment, e.g. `<ab n="0">`
- *Sentence*, as annotated by UD-Pipe, i.e. `<s>`

Positional attributes of vertical files:

- *word*: the token, e.g. `förbindelser`
- *lempos*: the lemma of the token with added part-of-speech, e.g. `förbindelse-n`;
- *tag*: the "PoS tag" of the token, given for convenience, e.g. `ADJ_NumSing_GenMasc_DegPos_Def_Nom`
- *pos*: the UD part-of-speech of the token, e.g. `ADJ`
- *feats*: UD features of the token, e.g. `Case=Nom | Definite=Def | Degree=Pos | Gender=Masc | Number=Sing`
- *deprel*: the UD dependency relation, e.g. `nmod`
- *head_word*, *head_lempos*, *head_tag*, *head_pos*, *head_feats*: same as above, but for the token that is the head of the current token
- *id*, *head_id*: the index of the token and its head

Corpus @ CLARIN.SI

Explore with KonText


kon text
Query Corpora Save Concordance Filter Frequency Collocations View Help

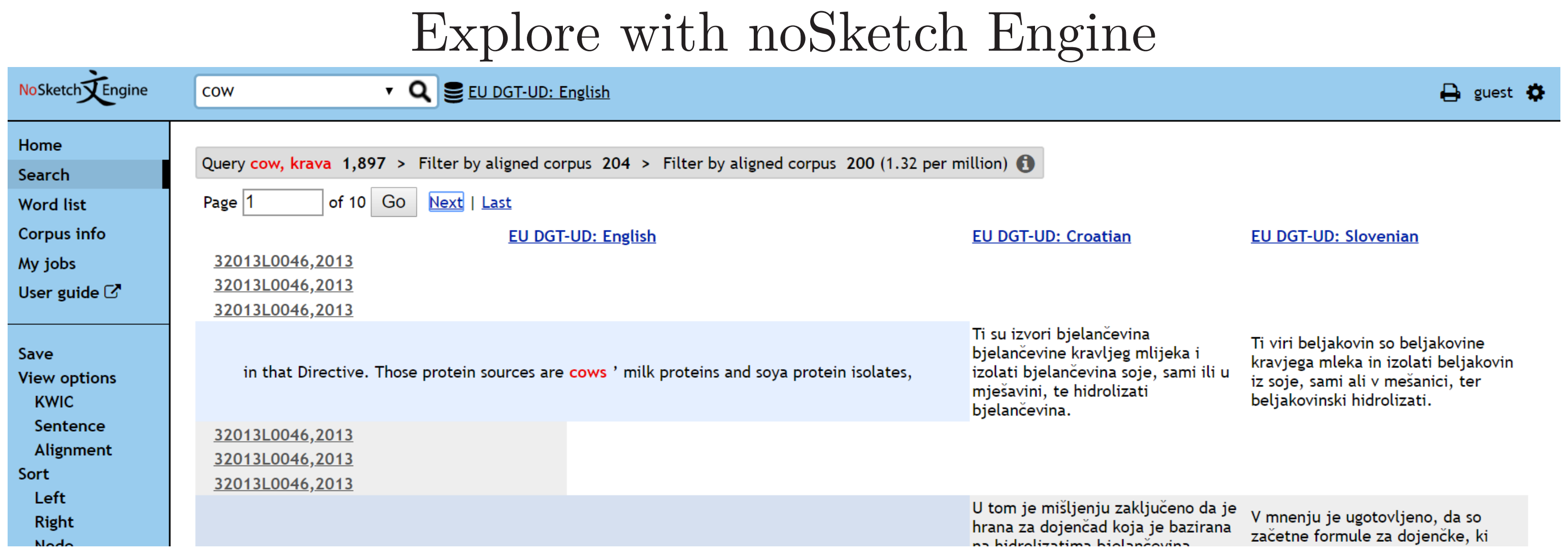
Corpus: EU DGT-UD: English | **Query:** ADJ, iron (129 hits)

Hits: **129** | i.p.m.: **0.85** (related to the whole "dgtud_en") | ARF: **24.44** | Result is sorted — 1 / 3

Line selection: simple ▾ | Display options

		EU DGT-UD: English		EU DGT-UD: Croatian		EU DGT-UD: Slovenian
□	32013D0448,2013	Accordingly, a benchmark has been developed for hot metal, with the product defined as **liquid**/ADJ_DegPos **iron**/NOUN_NumSing saturated with carbon for further processing.	32013D0448,2013	U skladu s time, određena je referentna vrijednost za vruću kovinu, a proizvod je definiran kao tekuće željezo zasićeno ugljikom za daljnju preradu.	32013D0448,2013	V skladu s tem je bila določena referenčna vrednost za vročo litino, ki je opisana kot tekoče železo, nasičeno z ogljikom za nadaljnjo predelavo.
□	32013L0059,2013	**Primary**/ADJ_DegPos **iron**/ADJ_DegPos production,	32013L0059,2013	Proizvodnja primarnog željeza,	32013L0059,2013	primarna proizvodnja železa;
□	32013R0936,2013	Manufacture of **basic**/ADJ_DegPos **iron**/NOUN_NumSing and steel and of ferro-alloys	32013R0936,2013	Proizvodnja sirovog željeza, čelika i ferolegura	32013R0936,2013	Proizvodnja surovega železa, jekla, ferozlitin
		Malleable/ADJ_DegPos **iron**/ADJ_DegPos castings for land		Dijelovi cestovnih vozila, klipnih motora i ostalih strojeva i		Deli za kopenska vozila, batne motorje ter druge

Explore with noSketch Engine

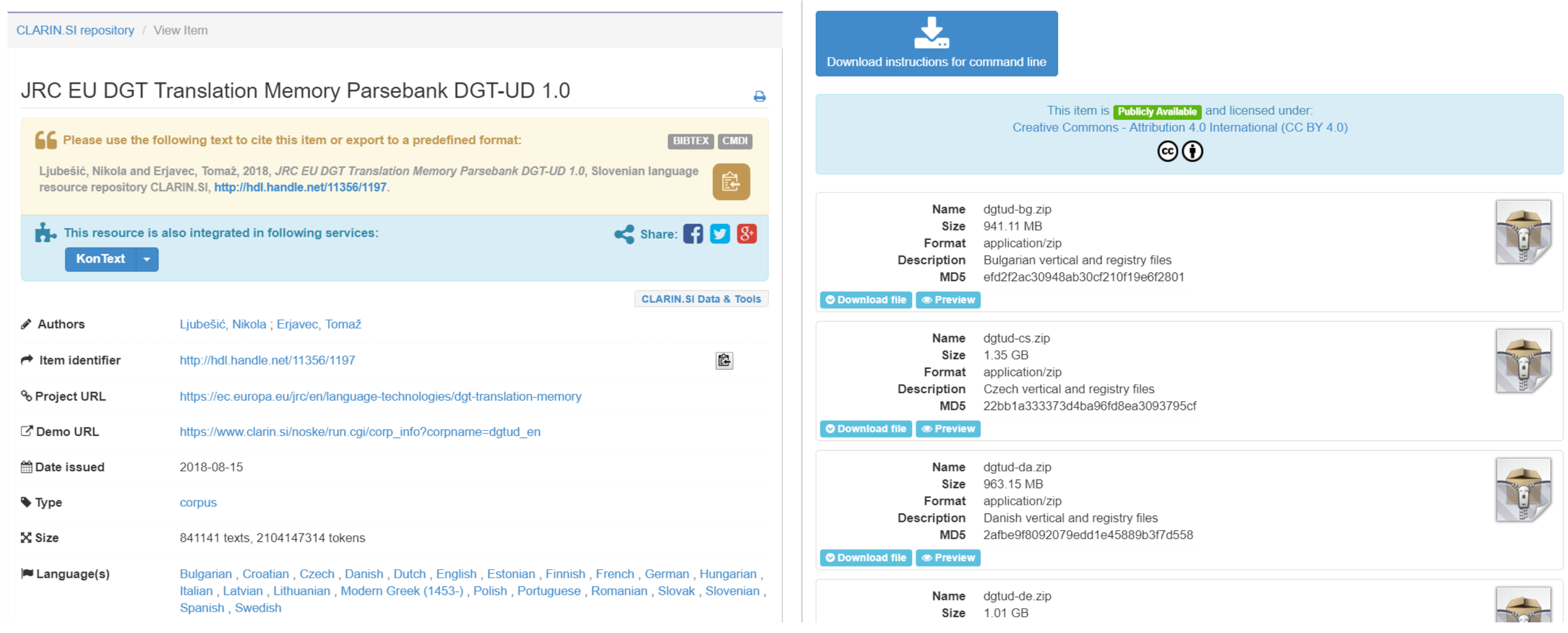


Query cow, krava 1,897 > Filter by aligned corpus 204 > Filter by aligned corpus 200 (1.32 per million)

Page 1 of 10 Go Next | Last

EU DGT-UD: English	EU DGT-UD: Croatian	EU DGT-UD: Slovenian
32013L0046,2013 32013L0046,2013 32013L0046,2013		
in that Directive. Those protein sources are **cows** ' milk proteins and soya protein isolates,	Ti su izvori bjelančevina bjelančevine kravljeg mlijeka i izolati bjelančevina soje, sami ili u mješavini, te hidrolizati bjelančevina.	Ti viri beljakovin so beljakovine kravjega mleka in izolati beljakovin iz soje, sami ali v mešanici, ter beljakovinski hidrolizati.
32013L0046,2013 32013L0046,2013 32013L0046,2013		
	U tom je mišljenju zaključeno da je hrana za dojenčad koja je bazirana	V mnenju je ugotovljeno, da so začetne formule za dojenčke, ki

Download corpus


CLARIN.SI repository / View Item
BIBTEX CMDI
Share:
KonText
CLARIN.SI Data & Tools
Download instructions for command line
Download file
Preview

JRC EU DGT Translation Memory Parsebank DGT-UD 1.0

Please use the following text to cite this item or export to a predefined format:

Ljubešić, Nikola and Erjavec, Tomaž, 2018, *JRC EU DGT Translation Memory Parsebank DGT-UD 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1197.

This resource is also integrated in following services:

Authors	Ljubešić, Nikola ; Erjavec, Tomaž
Item identifier	http://hdl.handle.net/11356/1197
Project URL	https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
Demo URL	https://www.clarin.si/noske/run.cgi/corp_info?corpname=dgtud_en
Date issued	2018-08-15
Type	corpus
Size	841141 texts, 2104147314 tokens
Language(s)	Bulgarian , Croatian , Czech , Danish , Dutch , English , Estonian , Finnish , French , German , Hungarian , Italian , Latvian , Lithuanian , Modern Greek (1453-) , Polish , Portuguese , Romanian , Slovak , Slovenian , Spanish , Swedish

This item is Publicly Available and licensed under:
Creative Commons - Attribution 4.0 International (CC BY 4.0)

Name	dgtud-bg.zip
Size	941.11 MB
Format	application/zip
Description	Bulgarian vertical and registry files
MD5	efd2f2ac30948ab30cf210f19e6f2801

Name	dgtud-cs.zip
Size	1.35 GB
Format	application/zip
Description	Czech vertical and registry files
MD5	22bb1a333373d4ba96fd8ea3093795cf

Name	dgtud-da.zip
Size	963.15 MB
Format	application/zip
Description	Danish vertical and registry files
MD5	2afbe9f8092079edd1e45889b3f7d558

Name	dgtud-de.zip
Size	1.01 GB