EUROPEAN NETWORK OF e-LEXICOGRAPHY

Between Corpora and Dictionaries / Crowdsourcing and Gamification
COST ENeL WG3 meeting
Budapest, Hungary, 24-25 February 2017

Extracting Lexical Data from Classification Schemes

Thierry Declerck, DFKI GmbH

Kseniya Egorova, Independent Researcher

Motivation

- We present current work dealing with a potential application in e-lexicography: the automatized creation of multilingual specialized dictionaries from structured data, which exists in the form of terminologies, classification schemes or taxonomies, etc.
- We implement an approach based on cross-languages and cross-taxonomies string mapping for automatically generating candidate multilingual dictionary entries for specific domains.

Data Sources

- We are currently testing this approach on two comparable multilingual industry classification schemes:
  - the Industry Classification Benchmark (ICB) http://www.icbenchmark.com
  - Global Industry Classification Standard (GICS) https://www.msci.com/gics

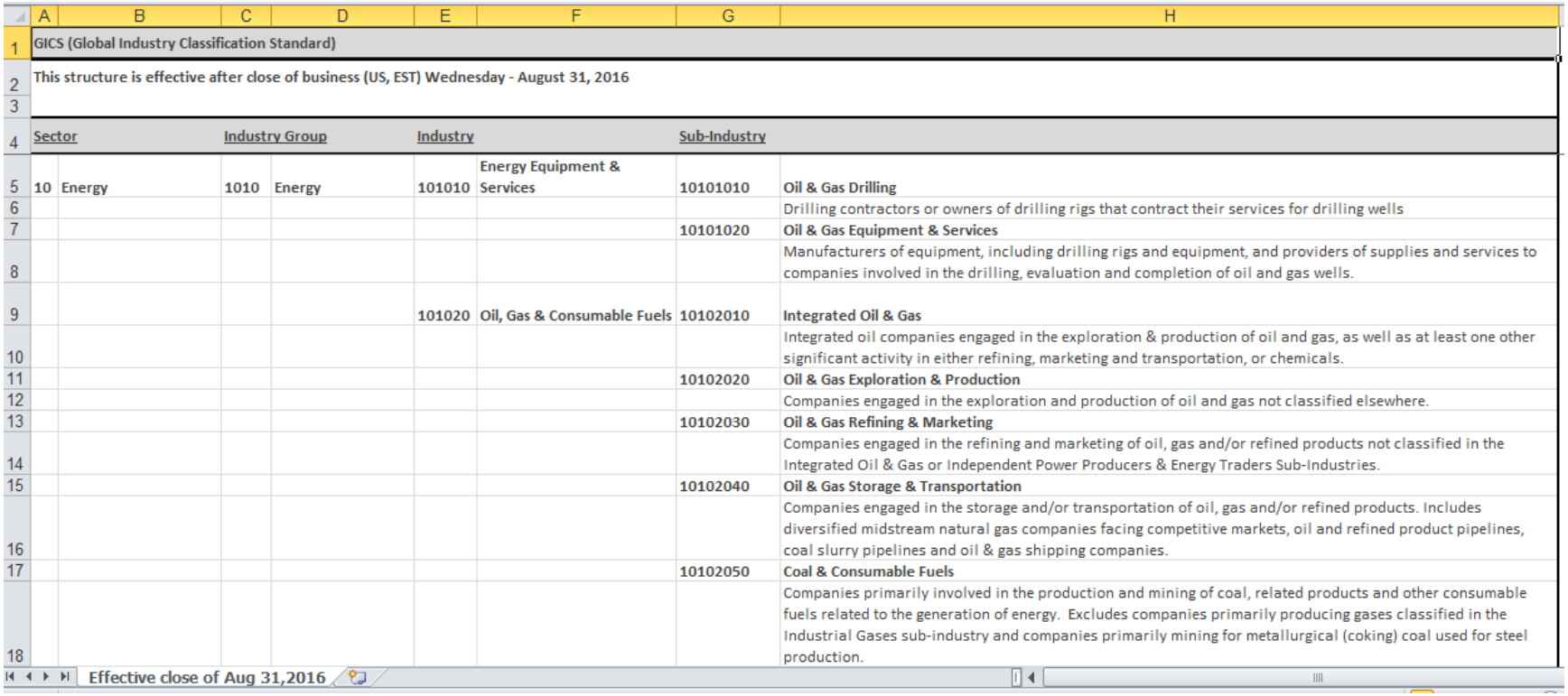
Structure of ICB

icb Industry Classification Benchmark

Industry Structure and Definitions

Industry	Supersector	Sector	Subsector	Definition
0001 Oil & Gas	0500 Oil & Gas	0530 Oil & Gas Producers	0533 Exploration & Production	Companies engaged in the exploration for and drilling, production, refining and supply of oil and gas products.
			0537 Integrated Oil & Gas	Integrated oil and gas companies engaged in the exploration for and drilling, production, refining, distribution and retail sales of oil and gas products.
		0570 Oil Equipment, Services & Distribution	0573 Oil Equipment & Services	Suppliers of equipment and services to oil fields and offshore platforms, such as drilling, exploration, seismic-information services and platform construction.
			0577 Pipelines	Operators of pipelines carrying oil, gas or other forms of fuel. Excludes pipeline operators that derive the majority of their revenues from direct sales to end users, which are classified under Gas Distribution.
		0580 Alternative Energy	0583 Renewable Energy Equipment	Companies that develop or manufacture renewable energy equipment utilizing sources such as solar, wind, tidal, geothermal, hydro and waves.
			0587 Alternative Fuels	Companies that produce alternative fuels such as ethanol, methanol, hydrogen and bio-fuels that are mainly used to power vehicles, and companies that are involved in the production of vehicle fuel cells and/or the development of alternative fuelling infrastructure.
1000 Basic Materials	1300 Chemicals	1350 Chemicals	1353 Commodity Chemicals	Producers and distributors of simple chemical products that are primarily used to formulate more complex chemicals or products, including plastics and rubber in their raw form, fiberglass and synthetic fibers.
			1357 Specialty Chemicals	Producers and distributors of finished chemicals for industries or end users, including dyes, cellular polymers, coatings, special plastics and other chemicals for specialized applications. Includes makers of colorings, flavors and fragrances, fertilizers, pesticides, chemicals used to make drugs, paint in its pigment form and glass in its unfinished form. Excludes producers of paint and glass products used for construction, which are classified under Building Materials & Fixtures.
	1700 Basic Resources	1730 Forestry & Paper	1733 Forestry	Owners and operators of timber tracts, forest tree nurseries and sawmills. Excludes providers of finished wood products such as wooden beams, which are classified under Building Materials & Fixtures.

Structure of GICS



	A	B	C	D	E	F	G	H
1	GICS (Global Industry Classification Standard)							
2	This structure is effective after close of business (US, EST) Wednesday - August 31, 2016							
3								
4	Sector		Industry Group		Industry		Sub-Industry	
5	10	Energy	1010	Energy	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling
6								Drilling contractors or owners of drilling rigs that contract their services for drilling wells
7							10101020	Oil & Gas Equipment & Services
8								Manufacturers of equipment, including drilling rigs and equipment, and providers of supplies and services to companies involved in the drilling, evaluation and completion of oil and gas wells.
9					101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas
10								Integrated oil companies engaged in the exploration & production of oil and gas, as well as at least one other significant activity in either refining, marketing and transportation, or chemicals.
11							10102020	Oil & Gas Exploration & Production
12								Companies engaged in the exploration and production of oil and gas not classified elsewhere.
13							10102030	Oil & Gas Refining & Marketing
14								Companies engaged in the refining and marketing of oil, gas and/or refined products not classified in the Integrated Oil & Gas or Independent Power Producers & Energy Traders Sub-Industries.
15							10102040	Oil & Gas Storage & Transportation
16								Companies engaged in the storage and/or transportation of oil, gas and/or refined products. Includes diversified midstream natural gas companies facing competitive markets, oil and refined product pipelines, coal slurry pipelines and oil & gas shipping companies.
17							10102050	Coal & Consumable Fuels
18								Companies primarily involved in the production and mining of coal, related products and other consumable fuels related to the generation of energy. Excludes companies primarily producing gases classified in the Industrial Gases sub-industry and companies primarily mining for metallurgical (coking) coal used for steel production.

Effective close of Aug 31,2016

ICB structures

Linking labels of categories in different languages

Four levels of classification, terminated by a definition of a company activities
The four levels are: Industry, Supersector, Sector, Subsector

ICB	**German Categories introducing the definition**	**English Categories introducing the definition**
Industry	7000 VERSORGER	7000 Utilities
Supersector	7500 Energieversorgung	7500 Utilities
Sector	7530 Elektrizität	7530 Electricity
Subsector	7537 Alternative Stromerzeugung	7537 Alternative Electricity

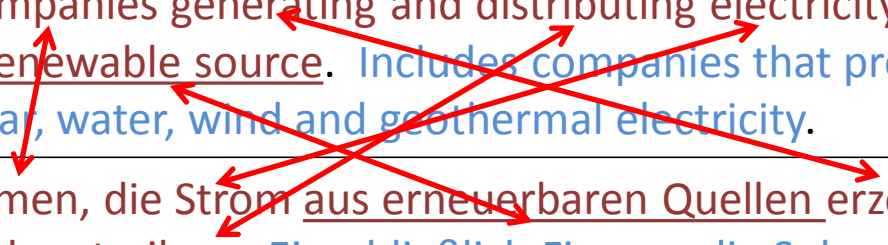
Observing that one and the same word in one language can be translated
by different words in the target language, depending on the level in the taxonomy:
Electricity => Elektrizität|Stromerzeugung;
Utilities => VERSORGER|Energieversorgung

ICB structures

Linking words used in definitions across language

Definitions associated with the subsector “7537 Alternative Electricity“

	7537 Alternative Electricity (EN) Alternative Stromerzeugung (DE)
English Definition	Companies generating and distributing electricity from a renewable source. Includes companies that produce solar, water, wind and geothermal electricity.
German Definition	Firmen, die Strom aus erneuerbaren Quellen erzeugen und vertreiben. Einschließlich Firmen, die Solar-, Wasser- und Windenergie sowie geothermische Energie erzeugen.



We observe: the term/word to be defined is probably the label “Alternative Electricity”. The intensional definition could be the first sentence in the definition, while the extensional definition could be in the second sentence But this might be valid only for this one example. Not sure if we can generalize. We can try to link the (sequence of) words in both languages.

Comparing to GICS

Like ICB: Also four levels of classification, terminated by a definition of a company activities
The four levels are: Sector, Industry Group, Industry, Sub-Industry

	German Categories introducing the definition	**English Categories introducing the definition**
Sector	55 Versorgungsbetriebe	55 Utilities
Industry Group	5510 Versorgungsbetriebe	5510 Utilities
Industry	551050 Unabhängige Energie- und Erneuerbare Elektrizitätshersteller	551050 Independent Power and Renewable Electricity Producers
Sub-Industry	55105020 Erneuerbare Elektrizität	55105020 Renewable Electricity

Comparing to GICS

Definitions in GICS are more extensive

	55105020 Renewable Electricity (EN) / Erneuerbare Elektrizität (DE)
English Definition	Companies that engage in the generation and distribution of electricity using renewable sources, including, but not limited to, companies that produce electricity using biomass, geothermal energy, solar energy, hydropower, and wind power. Excludes companies manufacturing capital equipment used to generate electricity using renewable sources, such as manufacturers of solar power systems, installers of photovoltaic cells, and companies involved in the provision of technology, components, and services mainly to this market.
German Definition	Unternehmen, die in der Herstellung und Verteilung von Elektrizität unter Verwendung von erneuerbaren Energien tätig sind. Eingeschlossen, aber nicht beschränkt auf, sind Unternehmen, die Elektrizität und/oder elektrische Leistung durch Energiequellen wie Biomasse, Biogase, Sonnen-, Wasser- und Windkraft herstellen. Ausgeschlossen sind Unternehmen, die in der Herstellung von Ausrüstungsgütern für Stromherstellung unter Verwendung erneuerbarer Energien tätig sind, wie Hersteller von Sonnenkraftanlagen, Installateure von photovoltaischen Zellen und Unternehmen, die ihre Technologie, Komponenten, und Dienste hauptsächlich diesem Markt anbieten.

We observe, like for ICB : the term/word to be defined is probably the label “Renewable Electricity ”. The intensional definition could be the first sentencein the definition, while the extensional definition could be in the second sentence. But this might be valid only for this one example. Not sure if we can generalize.
We can try to link the (sequence of) words in both languages.

Languages covered by GICS vs ICB

GICS: English and

- French
- German
- Italian
- Japanese
- Korean
- Portuguese
- Russian
- Simplified Chinese
- Traditional Chinese
- Spanish

ICB: English and

- Chinese
- Danish
- Estonian
- French
- Finnish
- German
- Icelandic
- Italian
- Japanese
- Latvian
- Lithuanian
- Spanish
- Swedish

Cross-Lingual Completions

- Idea: If we can connect DE in ICB and De in GICS for one category, we can then probably copy the Russian version of GICS into ICB and the other way round, the Estonian version of ICB into GICS
- At end, merging both taxonomies, at least for the labels

Adding the Russian definition for GICS 55105020 as a label for ICB 7537 Alternative Electricity (EN)

	7537 Alternative Electricity (EN) Alternative Stromerzeugung (DE)
English Definition	Companies generating and distributing electricity from a renewable source. Includes companies that produce solar, water, wind and geothermal electricity.
German Definition	Firmen, die Strom aus erneuerbaren Quellen erzeugen und vertreiben. Einschließlich Firmen, die Solar-, Wasser- und Windenergie sowie geothermische Energie erzeugen.
Russian Definition for GICS 55105020 (Renewable Electricity / Возобновляемые источники электроэнергии)	Компании, действующие в качестве производителей и поставщиков электроэнергии из возобновляемых источников. Включая, но не ограничиваясь, компании, которые производят электричество с использованием биомассы, геотермальной энергии, солнечной, гидроэнергии и энергии ветра. В данную группу не входят компании, производящие капитальное оборудование для выработки электроэнергии с использованием возобновляемых источников, такие как производители солнечных энергетических систем, компании, занимающиеся установкой фотоэлементов, в также компании, занимающиеся предоставлением технологий, компонентов и услуг в основном для этого рынка.

Adding the Estonian definition for ICB 7537 as a label for GICS 55105020 Alternative Electricity (EN)

	55105020 Renewable Electricity (EN) / Erneuerbare Elektrizität (DE)
English Definition	Companies that engage in the generation and distribution of electricity using renewable sources, including, but not limited to, companies that produce electricity using biomass, geothermal energy, solar energy, hydropower, and wind power. Excludes companies manufacturing capital equipment used to generate electricity using renewable sources, such as manufacturers of solar power systems, installers of photovoltaic cells, and companies involved in the provision of technology, components, and services mainly to this market.
German Definition	Unternehmen, die in der Herstellung und Verteilung von Elektrizität unter Verwendung von erneuerbaren Energien tätig sind. Eingeschlossen, aber nicht beschränkt auf, sind Unternehmen, die Elektrizität und/oder elektrische Leistung durch Energiequellen wie Biomasse, Biogase, Sonnen-, Wasser- und Windkraft herstellen. Ausgeschlossen sind Unternehmen, die in der Herstellung von Ausrüstungsgütern für Stromherstellung unter Verwendung erneuerbarer Energien tätig sind, wie Hersteller von Sonnenkraftanlagen, Installateure von photovoltaischen Zellen und Unternehmen, die ihre Technologie, Komponenten, und Dienste hauptsächlich diesem Markt anbieten.
Estonian definition for 7537 (Alternative Electricity / Alternatiivne elekter)	Ettevõtted, mis toodavad ja levitavad elektrit taastuvast allikast. Hõlmab ettevõtteid, mis toodavad päikese, vee, tuule ja maapõuenergiat. (correct linkings?)

Further Steps

- Need to lemmatize (at least certain) Wordforms for generating an entry in the domain specific dictionary.
- Linking to existing dictionaries and computational lexicons for obtaining all the possible surface forms (and see if one form is a privileged one in the domain specific context)
- Extract the definitions and compare with other taxonomic sources.
- Compare to generic uses of the extracted entries (for example with the Polyglot vectors, based on 117 language version of Wikipedia: http://polyglot.readthedocs.io/en/latest/Embeddings.html)

Other possible applications

- Cross-Lingual Ellipsis Resolution
  - Taking advantage of the use of hyphen compositiion in Germanic languages:
    - Erdöl- und Erdgasproduzenten => Erdölproduzenten und Erdgasproduzenten
    - Oil & Gas Producers => Oil Producers and Gas Producers
    - 石油･ガス精製 => 石油ス精製 ･ガス精製
    - Productores de petróleo y gas => Productores de petróleo de petróleo y Productores de gas

Other possible applications (2)

- Taxonomy extraction (from definition 7530):
  - Solar-, Wasser- und Windenergie” to Solarenergie, Wasserenergie and Windenergie. In the definition we also have the nominal phrase ‘geothermische Energie’
  - Can lead to: “Erneuerbare Quellen” delivers “Energie” and this type of “Energie” has various subtypes:
    - “Alternative Energie” hasSubclass “Solarenergie” ;
    - “Energie” hasSubclass “Windenergie” ;
    - “Energie” hasSubclass “Wasserenergie” ;
    - “Energie” hasSubclass “geothermische Energie” ;

Other possible applications (3)

- Ontology/Taxonomy merging/mapping
  - Where ICB and GICS have very similar definitions, labels and structure, can be merged.
  - Else: classes can be mapped to each other or complement each other
  - Merging and mapping with other relevant sources, like Gemet (GEneral Multilingual Environmental Thesaurus https://marinemetadata.org/references/gemet)
- Ontology/Taxonomic corrections: cross taxonomy and cross-lingual labels and definitions can lead to restructuration of the original sources.