



ИНИЦИАТИВИ НА ГОДИШНАТА КОНФЕРЕНЦИЯ НА КЛАРИН 2020

Първи ден, 5 октомври 2020

CLARIN CAFÉ: "THIS IS CLARIN. HOW CAN WE HELP YOU?"

КЛАРИН КАФЕ: КАК ДА ВИ ПОМОГНЕМ?

Steven Krauer представи какви са били мечтите и целите на CLARIN от началото, описа ролите на езика и колко полета за работа на CLARIN дават те. Езикът е носител на културно наследство, запис на миналото, основен инструмент за комуникация в и между обществата; езикът съхранява и разпространява знание; инструмент е за съставяне на правила в обществото; носител е на информация; той е начин за изразяване на хората; той е фокус на когнитивния процес; компонент на национална и културна идентичност; обект на изучаване; обект на компютърна обработка. Даде пример за това, което осигурява CLARIN: австрийски хуманитарист се консултира с CLARIN catalogue и намира подходящи текстови колекции в Швеция, Беларус и Саш; намира семантичен лексикон в Италия; инструмент за семантично аотиране в Германия. Използва ги, за да построи семантично аотирана версия на текстовите колекции с CLARIN interoperability standards. Проучва материала с помощта на център за експертиза в Испания. Слага аотирания текст в хранилище в България, за да се използва от други хора. Всичко това е направил от бюрото си.

Susanne Nylund Skog е (фолклорист от Швеция) описва богата колекция от аудиозаписи, която не представлява интерес за изследователите, ако не е транскрибирана. Транскрибирането на огромен архив отнема много време и е много трудоемко без подходящи инструменти. Говори за шведския проект ТилТол, в който обработват записите, за да ги направят достъпни за публиката. Даде пример със записи на известен фолклорист (Карл Нелсон) от миналото. Разполагат с общо 10 часа записи, в които той разказва истории и приказки, но често прекъсва една история, за да допълни нещо за друга и по този начин прави записите трудни за използване, защото автоматичното разпознаване на реч е напреднало, но все пак не може да се справи с всички проблеми все още. Целта на проекта им е да направят връзки между записите и свързаните с него документи – снимки, илюстрации, кореспонденция, други записи, но искат аудиозаписът да е в центъра на в Инструментите на CLARIN много им помагат. Спомена и многоезичен шведски проект за приказки от цял свят, които могат да се търсят по дигитална карта.

Christian Olesen представи проект за мултимедията на нидерландския CLARIN – правят колекции от снимки, текст, записи и т.н. за помощ на обучението. За да са полезни, колекциите разчитат на добри инструменти за текст и изображения. Работят по сложна колекция от неми филми, за да направят комплексен инструмент за анализ на мултимедия.

Noemi Vadasz (унгарски лингвист) – работи с анафори, кореференция. За докторантура е направила корпус с много слоеве за анотация и е направила модел за обработка на продрупове. Изтъкна ползите от инструментите и ресурсите на CLARIN, но и от събитията, които организира. Каза, че конференциите и уъркшопите на CLARIN дават възможност за изява на всички учени и тя лично е получила увереност от посещението на такива, защото е получила внимание, съвети, опит. Остави полезни линкове:

<https://www.clarin.eu/blog/e-magyar-comprehensive-processing-chain-hungarian>

<https://github.com/dlt-rilmta/emtsv>

<https://www.clarin.eu/blog/tour-de-clarin-hun-clarin-roadshows>

<https://www.clarin.eu/blog/natural-language-processing-historical-documents-%E2%80%93-workshop-report>

Darja Fiser попита какво биха посъветвали новодошлите.

Steven Krauer отовори, че е много важно къде се намираш – ако си от Германия и имаш въпроси, отиваш в съответния център на CLARIN. Ако си от държава, която не е член на CLARIN, трябва да се включваш в конференциите и други събития, за да се свържеш с хора от общността. Каза, че е важно да се информираме как други изследователи решават задачата, по която ние работим; че изследователите трябва да бъдат имитатори на добрите примери.

Noemi Vadasz отговори, че сайтът на CLARIN предлага много изчерпателна информация и всеки може да се запознае от там за събития и инструменти.

Ben Verhoeven изтъкна полезността на интервютата в сайта на CLARIN и попита какъв съвет биха дали на съществуващите центрове на CLARIN по света, как да помагат повече, как да се свързват по-лесно с хората.

Susanne Nylund Skog предложи да има гид във всеки център и отбеляза, че има страшно много ресурси, за които даже няма да как да разберем, че съществуват.

Ben Verhoeven каза, че е важно центровете да представят информация на достъпен за всички видове изследователи език.

Steven Krauer каза, че хората от различни дисциплини не говорят на един език и трябва да работят върху изграждане на общ език за комуникация. Попитаха **Steven Krauer** каква е мечтата му за CLARIN за бъдещето. Той каза, че е скучно, но много важно – трябва да се работи по авторските права, защото в момента можем да дигитализираме, но не и да споделяме. Каза също, че иска хората да работят в екип; че все още хуманитаристите работят индивидуално, а големите предизвикателства изискват големи екипи.

ЛИНКОВЕ ОТ ЧАТА:

- <https://www.maxkemman.nl/2019/01/collaborations-as-trading-zones/>
- <https://www.clarin.eu/blog/tour-de-clarin-clarin-pl-presents-websty-open-web-based-system-stylometric-analysis>
- <https://switchboard.clarin.eu/tools>
- <https://www.clarin.eu/content/parlamint>
- <https://www.clarin.eu/resource-families>

СЕСИЯ ЗА ЦЕНТРОВЕТЕ ЗА ЗНАНИЯ ЗА ЕЗИКА И ИЗСЛЕДВАНИЯ ПО ИЗКУСТВЕН ИНТЕЛЕКТ / (RESOURCES AND KNOWLEDGE CENTRES FOR LANGUAGE AND AI RESEARCH)

Jakob Lenard представи семействата от ресурси и инструменти на CLARIN. Към момента има 12 семейства корпуси, 5 семейства лексикални ресурси, 4 семейства инструменти.

Rachel Panckhurst представи достъпен онлайн корпус с SMS-и на френски.

Denis Eckert представи корпус с есперанто, направен от речници. Като проблем посочи, че не владеят всички езици от речниците, които ползват и търсят желаещи в CLARIN, за да направят малка общност от различни специалисти, с които да завършат задачата. Искат помощ с OCR. Разчитат, че CLARIN ще разпространи корпуса за аудиторията.

Dimitrios Kokkinakis представи проект за NLP и AI в помощ на невропсихологията – деменция, алцхаймер.

Eva Soroli представи френския knowledge център. Разказа за работните групи в центъра – практики и инструменти за корпусни изследвания, мултимодалност и нови форми на комуникация, многоезични корпуси, правни въпроси и защита на данни, добри практики за анотация на корпус и достъпност на корпусите. Чрез сайта си дават съвети за хранилища, стандарти, най-добри практики, методологии, стратегии. Най-често съветват хората да правят материалите си достъпни, да ползват стандарти, да използват TEI. Провеждат редовни обучения за докторанти и изследователи, за да дават информация за новите разработки.

Nikola Ljubesic представи CLASSLA – knowledge център за южни славянски езици. Каза, че приликите между езиците правят инструментите подходящи за много от тях. Каза, че онлайн уъркшопите, които организират, са поне три пъти по-посещавани от присъствените. Най-често задаваните въпроси към центъра са какви инструменти има и как да се използват. Към сайта на центъра са добавени *Често задавани въпроси* на няколко езика, за да се осигурява информация на потребителите. Имаше въпрос към Nikola Ljubesic - CLASSLA обхваща много задачи, това няма ли да накара правителствата на държави, които не са членове на CLARIN, да се откажат от включване в инфраструктурата? Никола отговори, че се надява точно на обратното - да се показва ползността на CLARIN, защото едва ли правителствата са запознати с нея и с ползите ѝ.

СЕСИЯ ЗА АНОТАЦИЯ И ВИЗУАЛИЗАЦИЯ (ANNOTATION AND VIZUALISATION)

Daniel de Kok представи Sticker2 - невронен синтактичен анотатор за нидерландски и немски, достъпен за всички в CLARIN. Основни са цели са били анотаторът да е бърз и лесен за ползване и да е standalone.

Richard Lawrence представи Rover - търсачка за синсети, направена за GermaNet. Каза, че е подобрение на стари ресурси. Направили са го на база на исканията на потребителите. Каза, че търсачката показва какво има в GermaNet и е много лесна за ползване - трябва ти само браузър. Имат търсене на регулярни изрази, търсене по част на речта и търсене на семантична свързаност.

Francesca Frontini показва ръчна анотация на Named Entities, предлагат и guidelines.

Jan Odijk представи полуавтоматичен анализ на реч на деца или хора с говорни дефекти. В центъра на задачата е граматичен анализ. Работят по автоматизация на процеса и предвиждат ползи за изследванията на говорните дефекти, насочват го към клиничните лингвисти.

Þórunn Arnardóttir представи невронен парсинг пайплайн за исландски, трениран на исторически текстове от няколко века. Имаше два почти еднакви въпроса към нея - защо са ползвали само исторически данни, щом обработват и съвременен исландски; какви са били резултатите на съвременния език. Тя отговори, че резултатите за съвременния исландски не са много добри и търсят корпус с повече данни, за да изследват език от подобни векове.

Подготвил: Зара Кънчева, ИИКТ-БАН