ИНИЦИАТИВИ НА ГОДИШНАТА КОНФЕРЕНЦИЯ НА КЛАРИН 2020

Втори ден, 6 октомври 2020

ПРЕЗЕНТАЦИЯ НА ПРЕДСЕДАТЕЛЯ НА ПРОГРАМНИЯ КОМИТЕТ COSTANZA NAVARETTA (PRESENTATION BY PROGRAMME COMMITTEE CHAIR

Costanza Navaretta - национален координатор на CLARIN за Дания - говори за подадените статии за конференцията тази година - само 40 статии, което е много малко и доказва, че кризата с COVID-19 се отразява на конференцията. Одобрените статии са само 36, но гарантират високо качество. Най-много автори са от Германия, Франция, Швеция, Естония. Статистиката показва, че 30 статии са на автори от една държава, а само шест статии - от различни. Този факт може да се тълкува като последица от COVID-19 кризата. Повечето статии са от един автор, съавторството е рядко. Бяха споменати и новостите в конференцията - панел за изкуствен интелект, студентска сесия, CLARIN in the classroom.

Изказа благодарности към целия екип, че успява да направи чудесна конференция в новата виртуална среда. Изтъкна като предимство на виртуалния формат факта, че няма паралелни сесии и участниците могат да посетят всички модули, както и, че по този начин всички развиваме уменията, необходими за онлайн събития. Каза, че CLARIN трябва да работи в тази посока - да осигури капацитет, за да помага на хората да организират виртуални формати, да подготвя експерти, защото те ще бъдат много необходими в бъдеще.

СЪСТОЯНИЕ НА CLARIN ИНФРАСТРУКТУРАТА (STATE OF THE CLARIN INFRASTRUCTURE)

Franciska de Jong показа, че инфраструктурата в момента има 24 члена и 3 центъра, които не са членове. Има 24 центъра от тип „Б“, а „К“ центровете нарастват най-бързо и покриват все повече теми.

Образуват се стратегически теми и екипи в CLARIN - иновативни методи за превод, обработка на хетерогенни данни и мултимедия, интегриране в облак на съществуващи инструменти и други. Изтъкна, че ресурсните семейства (https://www.clarin.eu/resource-families) са много полезни, помагат за колаборация между националните консорциуми; допринасят за повече сравнителни изследвания между различни езици и държави. Показа някои скорошни инициативи - курация на метаданни; ParlaMint, чиято цел е да построи множество от парламентарни корпуси, особено такива с публични дебати за кризи като COVID-19. Каза, че предвидените за пътувания средства на CLARIN ERIC, които не могат да се използват, осигуряват други възможности за финансиране (https://www.clarin.eu/funding). Каза, че трябва да се организират още повече уебинари и други виртуални събития. Говори за нуждата от нов дизайн на сайта на CLARIN, за да е още по-достъпен за потребителите - да се навигира по-лесно, да има нова секция - с лични истории на потребители и разработчици и показа видео с демо на сайта.

Dieter Van Uytvanck представи състоянието на техническата инфраструктура, промените от предходната година. Каза, че темата е важна, а често остава невидима за хората. Има два нови центъра „Б“, успешно преорганизиране на два стари центъра, четири други преорганизации текат в момента. Общо има 24 сертифицирани „Б“ центъра и 69 регистрирани. Показа статистика на влизането в сайта на CLARIN и новия опростен вариант за създаване на акаунт в него. Каза, че много усилия се полагат в посока на курацията на метаданни.

Говори за някои трудности, свързани с достъпността на информацията за CLARIN и публикациите, свързани с нея. Предизвикателства при търсенето на информация за структурата са испански писател и учен с фамилно име Clarin, които нямат връзка с CLARIN; индонезийска влогърка и община с това име; мебелна фирма и други.

Като сигурен източник за търсене на литература, свързана с CLARIN, изтъкна CLARIN Zotero (https://www.clarin.eu/zotero), който съдържа 3200 публикации. Каза, че се работи и в посока за създаване на постоянни идентификационни номера и собствени префикси за ограничаване на неяснотата и лесно разпознаване на цитати.В края на сесията се направи обща снимка на всички присъстващи, а по време на почивката се излъчваха видеа с интервюта на различни участници.

СЕСИЯ ЗА ХРАНИЛИЩА И РАБОТНИ ПРОЦЕСИ (REPOSITORIES AND WORKFLOWS)

Alois Pichler представи CLARINO+ - оптимизация на Wittgenstein Research Tools. Обясни, че това е един сет от данни, но с достъп по различни канали, за да може потребителите да избират най-удобния за тях.

Daniel de Kok представи работа по възпроизвеждане на анотационни ресурси за Weblicht. Каза, че са мотивирани от въпроса възможно ли е да възпроизвеждаме чужди научни данни. Целта е да интегрират Weblicht в други ресурси, да осигурят достъп на другите да ползват същото нещо по същия начин, но с възможността за промени; да осигурят софтуер, който може да се възпроизведе напълно. Изтъкна като сериозен проблем факта, че данните често имат нови версии, но моделите, които се тренират много продължително - не.

Paul Trilsbeek говори за работата по преместването на Езиковия архив на Института за психолингвистика „Макс Планк“ в хранилището на CLARIN, което е базирано на рамката на Islandora/Fedora. Отбеляза, че две години след промяната получават много добри отзиви от потребителите и администраторите, но и забелязват нуждата от малки модификации, които планират да завършат до 2022г.

Javier de la Rosa представи инфраструктура за анализ на поезия с фокус на онтологии и свързани отворени данни. Свободният достъп до корпуса е била една от основните цели на работата, авторът отбеляза, че много рядко се срещат отворени данни за поезия. Отбеляза, че една от трудностите е била липсата на стандарти в тази област - авторите са създали свои методи за управление на метаданни за хранилище с поезия. Получи въпрос дали моделът може да се прилага за други езици освен испански. Авторът отговори, че за момента не може, но лесно би могло да се пригоди за други романски езици; че в момента работят по езиков модел за многоезична обработка.

Maarten Janssen представи комбиниран инструмент за търсене от други два - KONTEXT (търсачка за езикови изследвания) и TEITOK (търсачка с визуализация). Полученият инструмент за търсене може да визуализира свързани ресурси - факсимилни страници, аудио и други.

Bart Jongejan представи преработка на стара система - Text Tonsorium, система за управление на работния процес за обработка на естествен език. „Tonsorium“ означава бръснарница, с което авторът подчертава, че ресурсът помага на потребителя бързо и лесно да настройва системата според нуждите си. Системата скоро ще бъде интегрирана в CLARIN DK (cst.dk).

СЕСИЯ ЗА КУРАЦИЯ НА ДАННИ, АРХИВИ И БИБЛИОТЕКИ (DATA CURATION, ARCHIVES AND LIBRARIES)

Niccolo Pretto говори за музикален архив - хетерогенна колекция, която може да се ползва от различни видове изследователи. Каза, че записите им са на касетки и имат лошо качество, затова използват много добър стандарт за подобряване на качеството при дигитализирането. Проектът им може да се разглежда като работа по запазване на стари данни.

При въпрос от **Maciej Piasecki** за видовете изисквания за качество на данните, Niccolo Pretto обясни, че е много трудно да имаш изисквания за формата - има много видове аналогово аудио; записите обикновено имат лошо качество, защото са записани с непрофесионална техника. Каза, че можеш да имаш много добър стандарт за достъп, запазване, архивиране, но е много трудно да наложиш критерии за всички общности.

Anne Ferger и **Daniel Jettka** представиха ISO стандарт за транскрибиране на реч. Казаха, че ползват различни техники за контрол на качеството и няколко мерки за изчислението му.

Aleksandr Riaposov представи QUEST - проект за създаване на критерии за качество и курация на анотиране на аудиовизуални езикови данни. Имат онлайн въпросник с два сценария за ползване - при планиране на работа (насочено към студенти и докторанти) и при приключване на проекти - с информация за достъпност, правни аспекти и други. Потребителите могат да качат данните си и да получат автоматична проверка на качеството. Каза, че качеството се измерва в това, дали данните са подходящи за две неща - архивиране и преизползване; дали съвпада със стандартите на общността.

Hanna Hedeland говори за липсата на широко приети стандарти за качество и документация на данните, което ги прави трудни за повторно използване. Каза, че често не можем да преизползваме данни именно заради качеството им.

Manfred Nolte представи проект за дигитализиране на университетски проекти и необходимостта от осигуряване на подходящи инструменти и ресурси за дигиталните хуманитаристи.

Federico Boschetti представи архив от средновековни латински текстове със свободен достъп в TEI формат, насочен към филолози, лингвисти и историци.

Laska Laskova представи проект за създаване на анотационна схема, която да обединява знания от български корпуси, речници, лингвистични анализатори с експертно знание от специалисти по история, диахронна лингвистика, иконография. Каза, че работата с различни експерти от консорциума включва обработка на много видове текст, за които е необходима унифицирана анотационна схема. Изтъкна важността на дискусиите за екипи от различни специалисти.

СЕСИЯ ЗА МЕТАДАННИ И ПРАВНИ АСПЕКТИ (METADATA AND LEGAL ASPECTS)

Aleksei Kelli представи рамка за споделяне на езикови данни. Каза, че споделянето на данни трябва да се разглежда като обработка на данни. Постави въпроса за отговорността при консорциумите - кой е отговорен, ако нещо се объркa? Университет? Институтът? Изследователите? Даде примери с различни подходи - в Естония отговорник и контрольор е университетът, във Франция - даден институт. Изрази мнение, че контролът и отговорността трябва да се дефинират от Европейския съюз. Каза, че GDPR е пример за глобално решение.

Vanessa Hannesschlager представи инструмент, който помага на изследователите да се запознаят с правните аспекти, свързани с работата и данните им. Каза, че изследователите не са адвокати и имат много трудности с разбирането на правните изисквания към изследванията си. Направили са колекции с правила като GDPR. Фокусът им е върху софтуерните лицензи, защитата на лични данни и помощта към широката публика на CLARIN.

Pawel Kamocki говори за авторските права върху н-грамите. Каза, че триграмите могат да се ползват свободно, седем- и десет-грамите са много трудни за използване.

На въпроса за отговорността при консорциумите отговори, че тя зависи от националния контекст и от институцията, въпреки че дефинициите за контрол са стриктни. Каза, че е важно как е организиран даден проект - от докторант или от асистент, защото е важно дали само обработваш данни или взимаш решения как те да бъдат обработвани и ползвани. Изрази вярване, че много скоро ще има общ европейски стандарт и закон за защита на данни

ПОЛЕЗНИ ЛИНКОВЕ:

- http://hdl.handle.net/11372/1030
- https://www.clarin.eu/reproducibility
- https://www.sshopencloud.eu/ssh-open-marketplace
- https://www.clarin.eu/news/stay-tuned-future-impact-research-infrastructures-social-sciences-and-humanities
- https://www.sshopencloud.eu/sshoc-train-trainer-bootcamp-librarians

Подготвил: Зара Кънчева, ИИКТ-БАН