Language technology solutions for lawyers - New tools and new perspectives workshop

Péter Homoki, lawyer

Pázmány Péter Catholics University
Faculty of Law and Political Sciences
Law and Language Research Group

14 February 2023

_Homoki
Ügyvédi
Iroda

Outline

Why small firms?

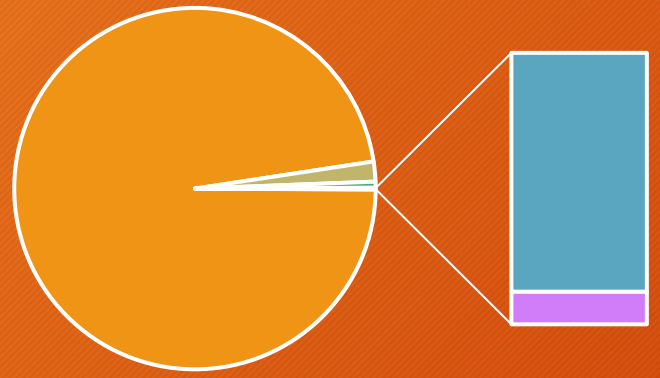
Small languages with small datasets in law

What applications?

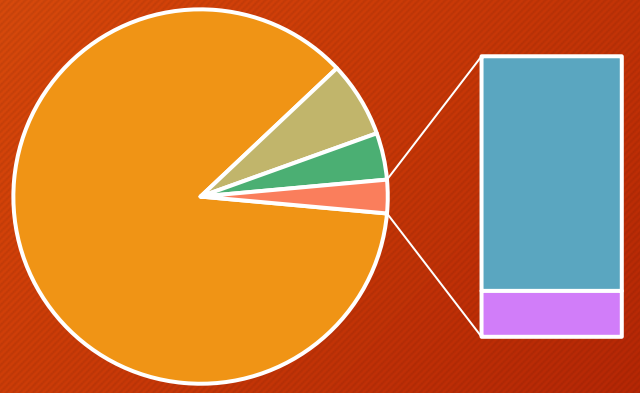
Changes to expect in the profession

https://ai4lawyers.eu/

EU

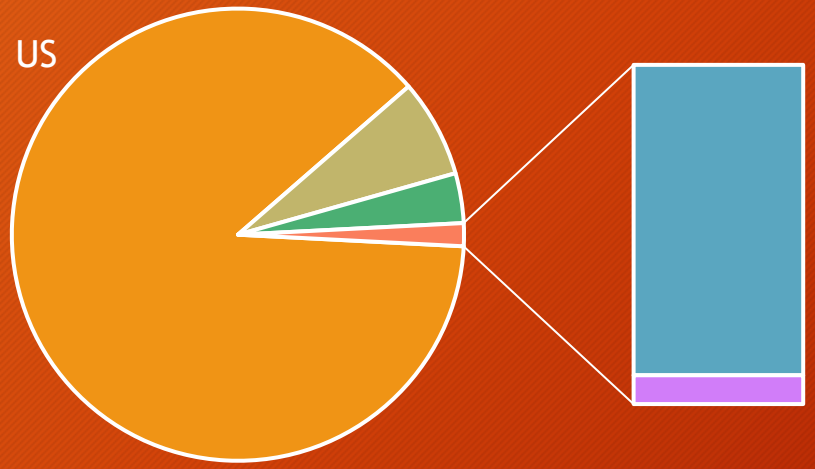


UK



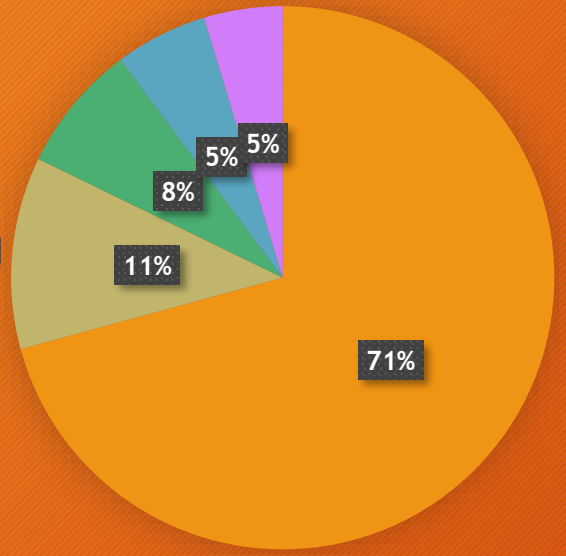
Establishments with 0 to 9 employees
Establishments with 10 to 19 employees
Establishments with 20 to 49 employees
Establishments with 50 to 249 employees
Establishments with 250 employees or more

US

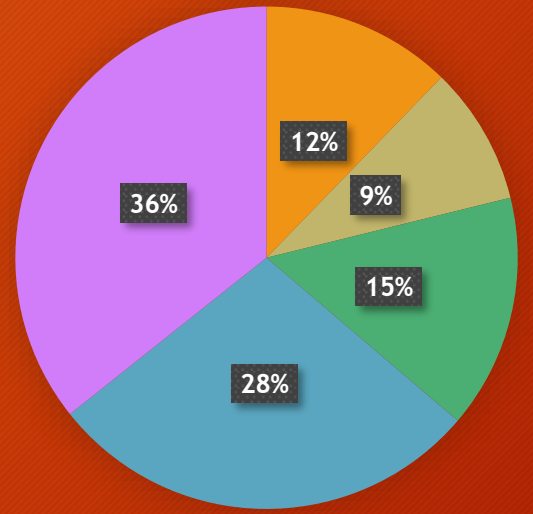


2018 % of establishment of all	EU establishment	US establishment	UK establishment
Entities with 0 to 9 employees	97.52%	87.80%	86.53%
Entities with 10 to 19 employees	1.78%	7.00%	6.54%
Entities with 20 to 49 employees	0.54%	3.56%	4.04%
Entities with 50 to 249 employees	0.13%	1.50%	2.41%
Entities with 250+ employees	0.02%	0.14%	0.47%

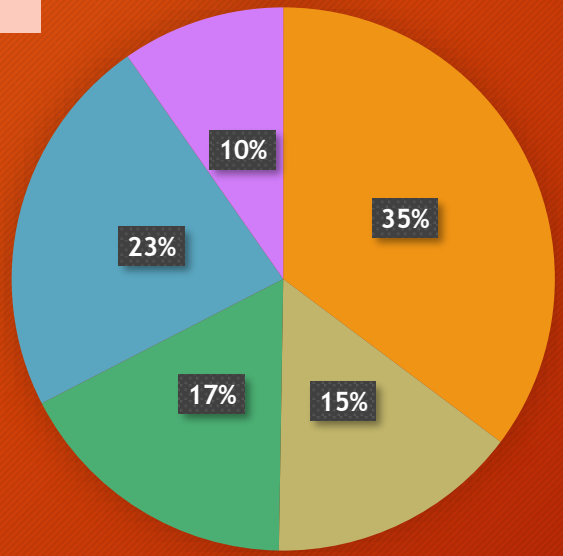
EU



UK



US



2018 % of employees	EU employees	US employees	UK employees
Entities with 0 to 9 employees	71.44%	35.23%	12.25%
Entities with 10 to 19 employees	11.54%	15.03%	8.91%
Entities with 20 to 49 employees	7.69%	17.18%	15.08%
Entities with 50 to 249 employees	5.56%	22.86%	28.07%
Entities with 250+ employees	4.72%	9.71%	35.69%

Small firms:

simple, flat organisation
tight budgets, lack of IT expertise, no developments
profit-oriented activity
strong competition within own market

Different business models for legal-like services

law firm business model: focusing on human, selling capacity of highly skilled individuals

vs.

legaltech based business model/legal operations

Small languages with small linguistic and legal datasets

- NLP tools/AI not magic: economies of scale needed for investment

- dataset needs of large language models: monolingual models

- excellent translation will not solve problems that need legal data to answer
- languages ≠ jurisdictions

- where may transfer learning help and where not?

WuDao trained on 1.2 TB Chinese, 1.2 TB English text
GPT-3 trained on 570 GB text

Model	Developer	Parameter Size
WuDao 2.0	Beijing Academy of Artificial Intelligence	1.75 trillion
MT-NLG	Nvidia and Microsoft	530 billion
Bloom	Hugging Face and BigScience	176 billion
GPT-3	OpenAI	175 billion
LaMDA	Google	137 billion
ESMFold	Meta AI	15 billion
Gato	DeepMind	1.18 billion

but training for legal applications
is not he same as training for
Large Language Models (LLM) from scratch;

legal models
may be built (finetuned) on LLMs

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

what applications?


Artificial Intelligence Index Report 2022
HAI Stanford University Human-Centered Artificial Intelligence
Computer Vision
Deep Learning
Facial Regonition
Knowledge Graphs
NL Generation
NL Speech Understanding
NL Text Understanding
Physical Robotics
Recommender Systems
Reinforcement Learning
Robotic Process Automation
Simulations
Transfer Learning
Virtual Agents
Industry
All Industries 23% 19% 11% 17% 12% 14% 24% 12% 17% 16% 26% 17% 12% 23%
Automotive and Assembly 15% 14% 9% 16% 3% 11% 12% 24% 12% 5% 33% 27% 6% 12%
Business, Legal, and Professional Services 29% 24% 15% 20% 23% 18% 19% 13% 22% 27% 31% 18% 21% 19%
Consumer Goods/Retail 23% 12% 14% 17% 11% 13% 14% 4% 8% 8% 16% 9% 1% 15%
Financial Services 17% 16% 11% 16% 12% 18% 32% 4% 13% 16% 33% 12% 12% 28%
Healthcare Systems/Pharma and Medical Products 30% 25% 12% 19% 10% 8% 26% 28% 22% 13% 28% 22% 19% 31%
High Tech/Telecom 28% 22% 6% 17% 17% 18% 34% 5% 19% 15% 23% 14% 11% 25%
% of Respondents (AI Capability)

Document generation: for speed of production and knowledge management

economies of scale needed for small firms?

Assisting drafting of documents: consistence, helping in quality rather than quantity

Legal research (NLU): access to larger databases, faster and wider research

Document analysis (NLU): faster and more precise research, specific reports of specific aspects, relying huge editorial type of work

RPA and help in internal office administration: relies on case and practice management systems, semi structured data already recorded etc., RPA not helpful for generic practices with no capacities to standardise workflow

transformation of small law firms?

loss of independence to tech. providers and intermediaries

increasing subscription/overhead or infrastructure costs

future of “PeopleLaw”? a future similar to dentists?
- forced changes in the working language
- more ADRs w/o lawyers instead of national litigation?
- role of new “internet” intermediaries

role of bars: tech providers, intermediaries or education?

Thank you!

_Homoki
Ügyvédi
Iroda

homoki.net