

Extracting Knowledge from Finnish Texts

Digital Humanities in Action: Sampo Model and Portals for Cultural Heritage, 29.10.2020

Minna Tamper

*Semantic Computing Research Group (SeCo), Aalto University, <https://seco.cs.aalto.fi>
minna.tamper@aalto.fi*

Content

- Use cases
- NLP Services
- Questions



Use cases

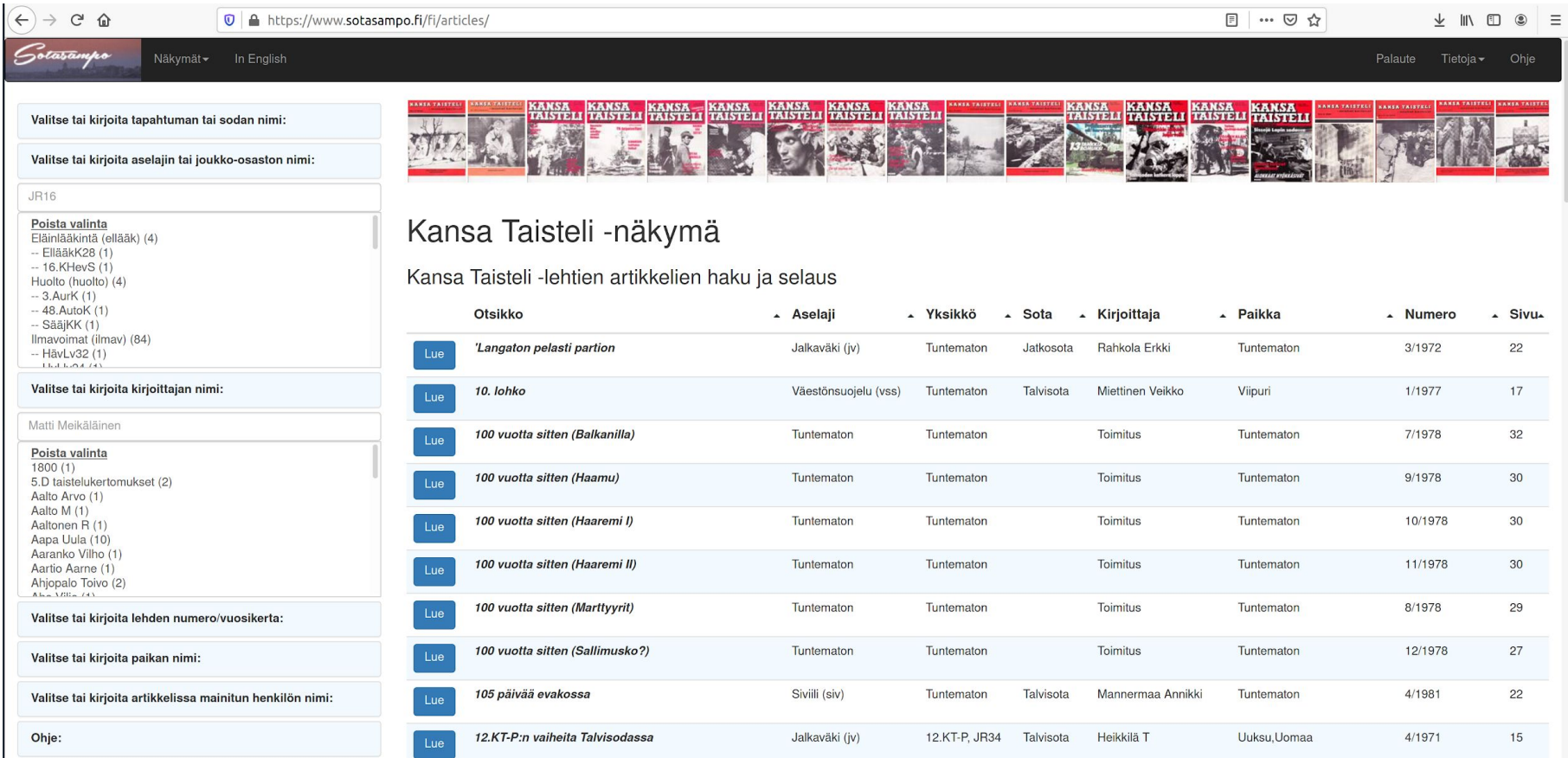


Aalto University
School of Science

Department of
Computer Science



WarSampo



BiographySampo

Linus Torvalds



* Helsinki 28.12.1969

filosofian kunniaatohtori, tietokoneohjelmointi

Lähteet: Kansallisbiografia. Kansallisbiografian sisällysluettelo

Biografiskt lexikon för Finland	Fennica	Geni.com: Kotisivu	Geni.com: Sukupu	Kotisivu	VIAF	Wikipedia	Wikidata
---	-------------------------	------------------------------------	----------------------------------	--------------------------	----------------------	---------------------------	--------------------------

[Kansallishistoria](#)
[Biografiakeskuksen artikkeli](#)
[Geni.com, kotisivu](#)
[Geni.com, sukupuu](#)
[Wikipedia](#)

KONTEKSTUAALINEN LUKIJA

Alla olevasta tekstistä on tunnistettu henkilöt, paikat, ja organisaatiot automaattisin menetelmin.

Linus Torvalds aloitti tietokoneiden Linux-käyttöjärjestelmän kehittämisen. Sitten tähän avoimeen ohjelmistokehitykseen ovat osallistuneet lukuisat vapaaehtoiset ohjelmoijat ympäri maailmaa, ja Torvaldsista on tullut uuden Internet-sukupolven kuulakuva ja tietotekniikka-alan kansainvälinen kuuluisuus.

Internet muuttii 1990-lulla maailmaa suunnaten sitä kohti tietokoneiden ja tietoverkkojen varassa toimivaa tietoyhteiskuntaa. Tämä maailmanlaajuinen muutos nosti tietotekniikan yhdeksi merkittävimmistä teollisuustuotannosta ja mahdollisesti samalla alavan uudenaikaista tietoverkkoihin perustuvia globaaleja tietoyhteisömuotoja. Kansainvälisenä vapaaehtoisten ohjelmajoina yhteistyönä kehitetty tietoyhteisöiden Linux-käyttöjärjestelmä sekä sen alullepanijasta [Linus Torvaldsista](#) uuden Internet-sukupolven keulakuvan ja tietokoneikka-alan kansainvälisen kuului

🟢 **Helsingissä** toimittajaperheessä kasva-
Törnqvistin kannustama tietokoneider
Helsingin yliopistossa tietojenkäsittelylle
ohjaavaan käyttöjärjestelmään Unixin. I-
hän halusi saada sellaisen omaan kotille
että Talvisu löydäntä markkinoilla sopiv-
toimivaa Unixa. Niinpä Törnqvist tallensi
kirjoittamaan ikkoma Unix-versiota tarka-

(1935-2013)



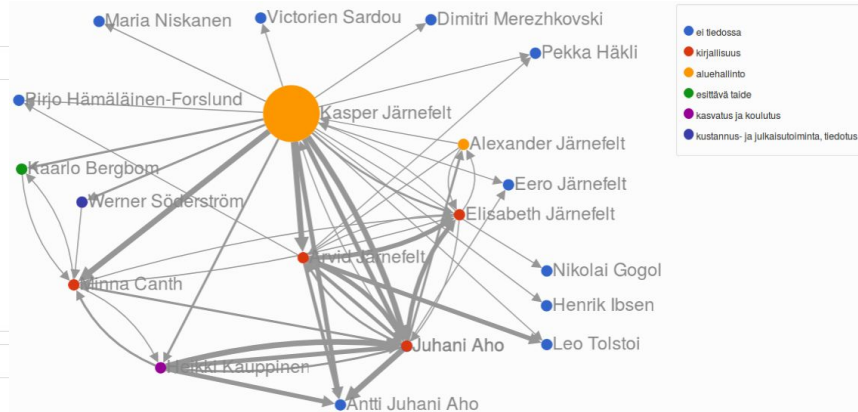
ana isoisänsä professori **Leo**
Lindus Törnqvist tutustui
tietokoneen toimintaan
järjestelmän siinä määrin, että
Ongelmana oli kuitenkin se,
vähän laadukasta kotitietokoneissa
atti 1991 ryhtyä harrastukseensa
in vain omaa käyttöönsä.

Matti Tienari

(1935-2013)



ana isoisänsä professori [Leo](#)
lut [Linus Torvalds](#) tutustui
in tietokoneen toimintoja
rjestelmään siinä määrin, että
. Ongelmana oli kuitenkin se,
vän laadukasta kotitietokoneissa
ätti 1991 ryhtyä harrastukseensa
in vain omaan käyttöönsä.



Verb	All	Noun	All	Adjective	All	Proper noun	All	Verb	All	Noun	All	Adjective	All	Proper noun	All
be	1.90%	year	0.64%	political	2.68%	Finnish	1.00%	be	1.59%	year	0.74%	first	0.31%	Finnish	0.92%
become	0.34%	person	0.17%	human	0.15%	Swedish	0.15%	become	0.34%	year	0.17%	first	0.17%	Finnish	0.17%
get	0.33%	government	0.57%	first	0.61%	Kekkonen	0.21%	become	0.33%	chairman	0.05%	Finnish	0.22%	Turku	0.13%
act	0.29%	party	0.49%	Finnish	0.17%	SDP	0.25%	act	0.30%	member	0.49%	the	0.17%	SDP	0.12%
have	0.28%	time	0.44%	political	0.44%	Yrjöng	0.12%	have	0.28%	time	0.11%	new	0.11%	the	0.11%
keep	0.16%	time	0.34%	own	0.31%	Soviet Union	0.12%	keep	0.16%	government	0.40%	large	0.15%	Hälsä	0.10%
make	0.15%	Congressman	0.30%	second	0.12%	Uola	0.09%	make	0.15%	Congressman	0.30%	Socialist	0.13%	bridgehead	0.09%
lead	0.14%	time	0.29%	political	0.23%	Yrjöng	0.12%	lead	0.14%	time	0.11%	new	0.11%	the	0.11%
to lead	0.13%	union	0.20%	key	0.08%	Finnish	0.08%	to lead	0.13%	time	0.31%	second	0.11%	Mary	0.09%
be	0.12%	time	0.18%	political	0.18%	Yrjöng	0.12%	be	0.12%	time	0.11%	new	0.11%	the	0.11%
establish	0.11%	mission	0.18%	significant	0.07%	Pietari	0.07%	establish	0.11%	job	0.20%	next	0.10%	Helsing	0.08%
participate	0.11%	president	0.17%	young	0.07%	Uola	0.07%	participate	0.11%	job	0.20%	size	0.09%	Mire	0.08%
be	0.10%	time	0.17%	general	0.07%	Yrjöng	0.07%	be	0.10%	articles	0.10%	Germany	0.07%	the	0.07%
start	0.10%	management	0.16%	society	0.07%	Paasikivi	0.06%	start	0.10%	vice chairman	0.11%	good	0.07%	Hälsä	0.07%
to move	0.09%	minister	0.14%	Socialist	0.06%	Linnela	0.06%	to move	0.09%	family	0.16%	young	0.07%	Wojtyła	0.07%
be	0.09%	time	0.14%	may	0.06%	Yrjöng	0.06%	be	0.09%	landmark	0.10%	the	0.07%	the	0.07%
influence	0.09%	universally	0.04%	Democratic	0.06%	Karelia	0.05%	influence	0.09%	mission	0.16%	important	0.07%	Isä	0.07%
be	0.08%	time	0.13%	general	0.06%	Peikari	0.05%	be	0.08%	to lead	0.08%	Socialist	0.07%	Soviet Union	0.07%
be	0.08%	Board	0.07%	own	0.06%	Kuusinen	0.05%	be	0.08%	policy	0.13%	Democratic	0.07%	Hagman	0.06%
be	0.08%	international	0.13%	general	0.06%	Uola	0.05%	be	0.08%	policy	0.13%	Democratic	0.07%	Hagman	0.06%
start	0.07%	policy	0.13%	case	0.06%	Uusmaa	0.05%	start	0.07%	home	0.12%	Swedish	0.06%	Alno	0.06%
write	0.07%	manager	0.12%	coldest	0.06%	Westerlund	0.05%	write	0.07%	president	0.12%	Communist	0.06%	Parliament	0.06%
be	0.07%	time	0.12%	new	0.06%	Chapman	0.05%	be	0.07%	to move	0.07%	the	0.06%	the	0.06%
get	0.07%	Agrarian	0.16%	many	0.06%	Kuopio	0.05%	get	0.07%	daughter	0.12%	active	0.06%	Alno	0.06%
start	0.07%	time	0.12%	knows	0.06%	Uola	0.05%	start	0.07%	vista	0.12%	general	0.06%	the	0.06%
continue	0.07%	job	0.11%	BCD	0.06%	Brother	0.05%	continue	0.07%	brother	0.12%	young	0.06%	Hälsä	0.06%
return	0.06%	child	0.11%	Communist	0.06%	Europe	0.05%	return	0.06%	life	0.12%	test	0.05%	Tekla	0.05%
manage	0.06%	time	0.11%	Yrjöng	0.06%	Yrjöng	0.05%	manage	0.06%	to lead	0.06%	the	0.05%	the	0.05%
represent	0.06%	inhabitant	0.11%	difficult	0.04%	Sweden	0.04%	represent	0.06%	Supervisory	0.11%	strong	0.05%	Yrjöng	0.05%
arrive	0.06%	time	0.11%	active	0.04%	Russia	0.04%	arrive	0.06%	Board	0.08%	different	0.04%	Hälsä	0.05%
be	0.06%	secondary	0.04%	time	0.04%	Sweden	0.04%	be	0.06%	board	0.11%	Swedish	0.04%	Hälsä	0.05%
So no	0.06%	school	0.11%	active	0.04%	Russia	0.04%	So no	0.06%	career	0.11%	small	0.04%	Hälsä	0.05%
constituted	0.06%	graduated	0.10%	wide	0.04%	Uola	0.04%	constituted	0.06%	thing	0.10%	wide	0.04%	Uusimaa	0.05%
after	0.05%	after	0.10%	Kokkari	0.04%	Kokkari	0.04%	after	0.05%	other	0.10%	the	0.04%	Kokkari	0.05%
leave	0.05%	celius	0.10%	test	0.04%	Estimann	0.04%	leave	0.05%	experience	0.05%	public	0.04%	Kylläski	0.05%
feel	0.05%	source	0.10%	strong	0.04%	Uola	0.04%	feel	0.05%	source	0.10%	the	0.04%	the	0.04%
feel	0.05%	source	0.10%	strong	0.04%	Uola	0.04%	feel	0.05%	history	0.10%	southern	0.04%	Vers	0.05%
present	0.05%	station	0.10%	high	0.03%	Mauno	0.04%	present	0.05%	look	0.05%	economic	0.04%	Koivisto	0.05%
accept	0.05%	social democratic	0.10%	high	0.03%	Uola	0.04%	accept	0.05%	work	0.10%	the	0.04%	Koivisto	0.05%
accept	0.05%	representative	0.09%	third	0.03%	Erkki	0.04%	accept	0.05%	school	0.09%	green	0.03%	Välimäki	0.05%
look	0.05%	senator	0.09%	intermediate	0.03%	Ants	0.03%	look	0.05%	work	0.09%	private	0.03%	Americ	0.05%
demand	0.05%	demand	0.09%	work	0.03%	Uola	0.04%	demand	0.05%	school	0.09%	the	0.03%	the	0.03%

LawSampo

Results

Legend: person, animal, mythical or fictional person,

general location, address, political location (e.g. state), geographical location, buildings or structures, astronomical locations (e.g. planets, galaxies),

organization, media organization, financial organizations, corporation and administration,

date, time, product, event, units (e.g. grams, meters), money, registry numbers, social security numbers, statutes, case law, domain information, title, and unknown entity ?.

" [Työntekijän](#)⁵ palkkasaatavia koskeva [kanne](#)³ oli jätetty tutkimatta, koska sitä ei ollut nostettu [työsopimuslain 13 luvun 9 §](#)¹:n 3 momentissa säädettyssä kahden vuoden [määräajassa](#)⁴ [työsuhteen](#)² päättymisestä vaan vasta [3,5](#)⁰ vuoden kuluttua siitä. Saman säännöksen mukaan palkkasaatava vanhentuu kuitenkin pykälän 1 momentissa säädetyn tavoin ja siis viiden vuoden kuluttua [erääntymispäivästä](#)⁸, jos [työntekijän](#)⁵ saatavan perusteena olevia [työehtosopimuksen](#)⁷ [määräyksiä](#)⁶ on pidettävä ilmeisen tulkinnanvaraisina. [Kanteessa](#)¹⁰ oli [kysymys](#)⁹ [työsuhteeseen](#)² sovellettavasta [työehtosopimuksesta](#)⁷ ja tulkinnanvaraisesta työehtosopimuksen soveltamisalaa koskevasta [määräyksestä](#)⁶. Myös sellaista voitiin pitää palkkasaatavan perusteena olevana ilmeisen tulkinnanvaraisena [määräyksenä](#)⁶. [Kannetta](#)¹⁰ ei olisi saanut jättää 3 momentin nojalla vanhentuneena tutkimatta. [TSL 13 luku 9 §](#)¹¹ 3 mom

"

Time

0. 3,5

Statutes, Court decisions

1. työsopimuslaki 13 luku 9 §

11. TSL 13 luku 9 §

Related documents:

[ECLI:FI:KKO:2011:85](#)

[ECLI:FI:KKO:2016:19](#)

[ECLI:FI:KHO:2001:T894](#)

[ECLI:FI:KHO:2001:29](#)

[ECLI:FI:KHO:1987:T1261](#)

[ECLI:FI:KKO:1980:II77](#)

JSON

Download JSON

NLP Services



Aalto University
School of Science

Department of
Computer Science



Other datasets and services

- **Henko - person name ontology**
 - Based on Finnish Digital Agency's name data and person names collected from other Sampos.
- **Gender identifier**
 - Tool for making a sophisticated guess about person's gender based on Henko's statistics of name usage
- **Person name identifier**
 - Tool to aid in extraction of person names from text, enriching them with other contextual information

Platform for NLP Services

- Publishing APIs for NLP processing
 - <https://nlp.ldf.fi>

SeCo Text Annotation Service is a set of different tools that can be used for Finnish text and language analysis. The tools exchange text to JSON that contains results of each tool.

Select a tool of choice

FINER

Named entity recognition based on the FIN-CLARIN's FINER tool.

FinBERT

TurkuNLP's named entity recognition application based on Google's BERT.

APPI - Automatic annotation tool

Named entity linking based on voting scheme.

Finnish dependency parser

Linguistic analysis based on the Turku NLP Finnish-dep-parser

Date and registry number identifiers

Regular expression based application for identifying registry numbers, dates, and more.

Person name finder service

Application that can pick up person names from text and offers links and contextual information (gender, lifespan) about each name.

Gender Identification Service

Application that can determine person's gender based on a name.



Aalto University
School of Science

Department of
Computer Science



Next

- **Service for transforming text to RDF**
- **Applying NLP methods and processes in LawSampo, ParliamentSampo**
- **Application for searching, classifying and exploring legal documents based on semantic content**
- **Other applications coming**

More information

- **Homepage:**
 - <https://seco.cs.aalto.fi/projects/nlp/>
- **EU COST Action**
 - <https://nexuslinguarum.eu/>



Questions?



Aalto University
School of Science

Department of
Computer Science

