Evaluating Trends in Czech and English

Ondřej Herman

Trends

- Identify words increasing or decreasing in usage
- Linear trend
  - Linear regression
  - Robust statistics (Theil-Sen slope, Mann-Kendall p-value)
- Implemented in Sketch Engine
- Quality of the result unclear

TRENDS

English Trends (2014–today)

Get more space

#	Lemma	Trend ↓	Frequency	Sample
1	ultrawide	↗ 3.27	63,488	
2	unhoused	↗ 3.27	53,617	
3	multiomic	↗ 3.27	1,396	
4	minoritised	↗ 3.27	1,594	
5	reshared	↗ 3.08	11,838	
6	gaslit	↗ 3.08	7,015	
7	neurodiverse	↗ 3.08	12,526	
8	deeptech	↗ 3.08	3,337	
9	agrivoltaic	↗ 3.08	2,952	
10	chonky	↗ 3.08	2,353	
11	cottagecore	↗ 3.08	5,779	
12	multiomics	↗ 3.08	1,321	
13	altcoin	↗ 2.90	102,199	
14	chiplet	↗ 2.90	15,047	
15	minoritized	↗ 2.90	4,559	
16	parasocial	↗ 2.90	5,505	
17	substack	↗ 2.90	3,074	
18	explainability	↗ 2.90	10,127	
19	hyperscaler	↗ 2.90	18,826	
20	hyperpop	↗ 2.90	3,804	

Rows per page: 20 1–20 of 1,000 1 / 50

TRENDS

English Trends (2014–today)

Get more space

	Lemma	Trend ↓	Frequency	Sample
1	ultrawide	↗ 3.27	63,488	•••
2	unhoused	↗ 3.27	53,617	•••
3	multiomic	↗ 3.27	1,396	•••
4	minoritised	↗ 3.27	1,594	•••
5	reshared	↗ 3.08	11,838	•••
6	gaslit	↗ 3.08	7,015	•••
7	neurodiverse	↗ 3.08	12,526	•••
8	deeptech	↗ 3.08	3,337	•••
9	agrivoltaic	↗ 3.08	2,952	•••
10	chonky	↗ 3.08	2,353	•••

	Lemma	Trend ↓	Frequency	Sample
11	cottagecore	↗ 3.08	5,779	•••
12	multiomics	↗ 3.08	1,321	•••
13	altcoin	↗ 2.90	102,199	•••
14	chiplet	↗ 2.90	15,047	•••
15	minoritized	↗ 2.90	4,559	•••
16	parasocial	↗ 2.90	5,505	•••
17	substack	↗ 2.90	3,074	•••
18	explainability	↗ 2.90	10,127	•••
19	hyperscaler	↗ 2.90	18,826	•••
20	hyperpop	↗ 2.90	3,804	•••

Rows per page: 20 1–20 of 1,000 1 / 50

Test Data: Trend Corpora

- Corpora built by Lexical Computing from RSS feeds
- 25+ languages, updated twice a week with fresh timestamped data

- October 2024 – September 2025 (1 year period)
- March 2023 – February 2025 (2 year period)

X

- Czech
- English

Evaluation

- One human annotator
  - Examined the words using Sketch Engine
- GPT-OSS-120B
  - 25 random concordance lines for each word
  - Frequency list for the past 5 years to allow seasonality detection

Evaluation

- Take top 100 trending lemmas according to Mann-Kendall / Theil-Sen
- Lowercase only
- Classify them as one of the following:

1. **Proper Name** - not interesting from linguistic point of view
2. **Seasonal Word Usage** - is trending periodically, not specific
3. **Processing Error** - meaningless / too mangled to be useful, not a word
4. **Source Composition Issue** - comes from a single website
5. **OK** - no other issue
6. **Interesting Trending Word** - same as above, but looks interesting to me

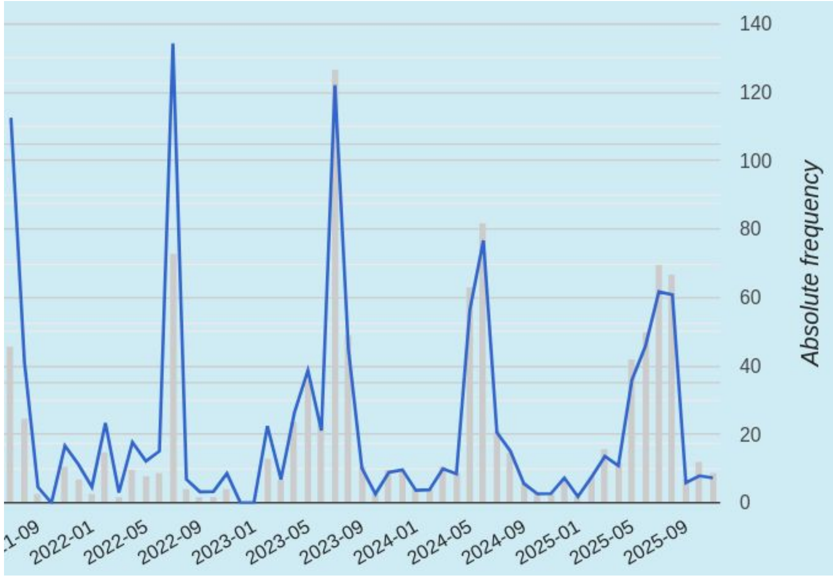
Evaluation

1. **Proper Name** - not interesting from linguistic point of view
    - Czech: al-Udeid, plánička
    - English: duPont, i7-14700F

Evaluation

2. Seasonal Word Usage - is trending periodically, not specific

- Czech: mandelinka, cyklovýleta
- English: quarter-time, lifeguarded



Evaluation

3. Processing Error - meaningless / too mangled to be useful / not a word

- Czech: homeAktualit, =============
- English: definitelyNo, =+04:00-

Evaluation

4. Source Composition Issue - only appears on a single website

- Czech: kategorizační
- English: deworming, xwOBA

Evaluation

5. OK - no other issue

- Czech: krémovost, termála
- English: digitoxin, microschool

6. Interesting Trending Word - as above, but looks subjectively interesting

- Czech: greenfluencerka, samorozvrhování
- English: open-weight, romantasy

Category	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
	Human	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

	Czech				English			
	2-year period		1-year period		2-year period		1-year period	
Category	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**	**Human**	**LLM**
OK (Interesting)	**59** (7)	**39** (39)	**57** (6)	**31** (29)	**54** (5)	**23** (18)	**47** (2)	**33** (30)
Proper Name	22	22	10	8	15	17	13	13
Seasonal Word Usage	0	8	24	45	0	2	9	15
Processing Error	18	16	9	8	27	36	22	19
Source Composition Issue	1	15	0	8	4	22	9	20
Cohen's κ	0.436		0.400		0.288		0.602	

Evaluation Issues

- The classification is fuzzy
- The categories overlap
  - e.g. **Processing Error** commonly implies **Source Composition Issue**
- How much diversity in sources should we require?
- More annotators?

Conclusion

- Approximately 30 % - 50 % of trending words seem to be salient.
- LLM and a human annotator agree reasonably, Cohen’s κ ~ .3 - .6
    - I don’t trust either of them.
- LLM filtering of the result seems to be useful.