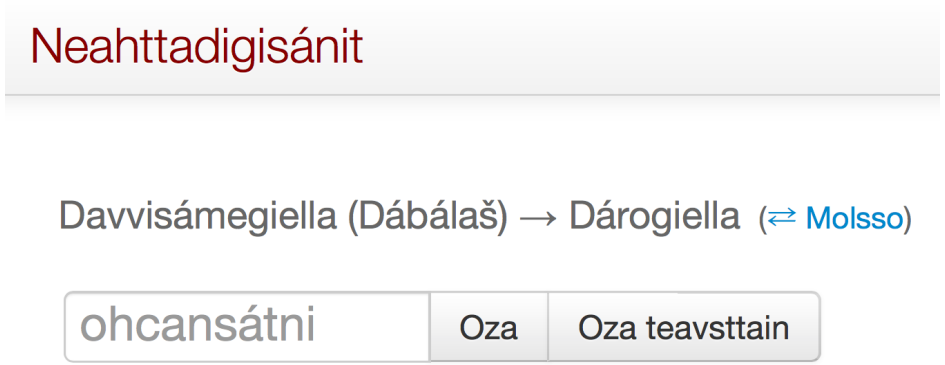
What can the online dictionary log data tell us?

Trond Trosterud, Lene Antonsen and Ciprian Gerstenberger
Giellatekno
UiT Arctic university of Norway



Umeå, March 10th 2016

Overview

Imagine just clicking in the text to get a translation

- But for Saami you cannot just do that
- The chance of hitting a word in its dictionary form is low (Antonsen et al 2009):
  - North Saami: 7,9 %
  - Finnish: 10,0 %
  - Norwegian 30,5 %

Imagine just clicking in the text to get a translation

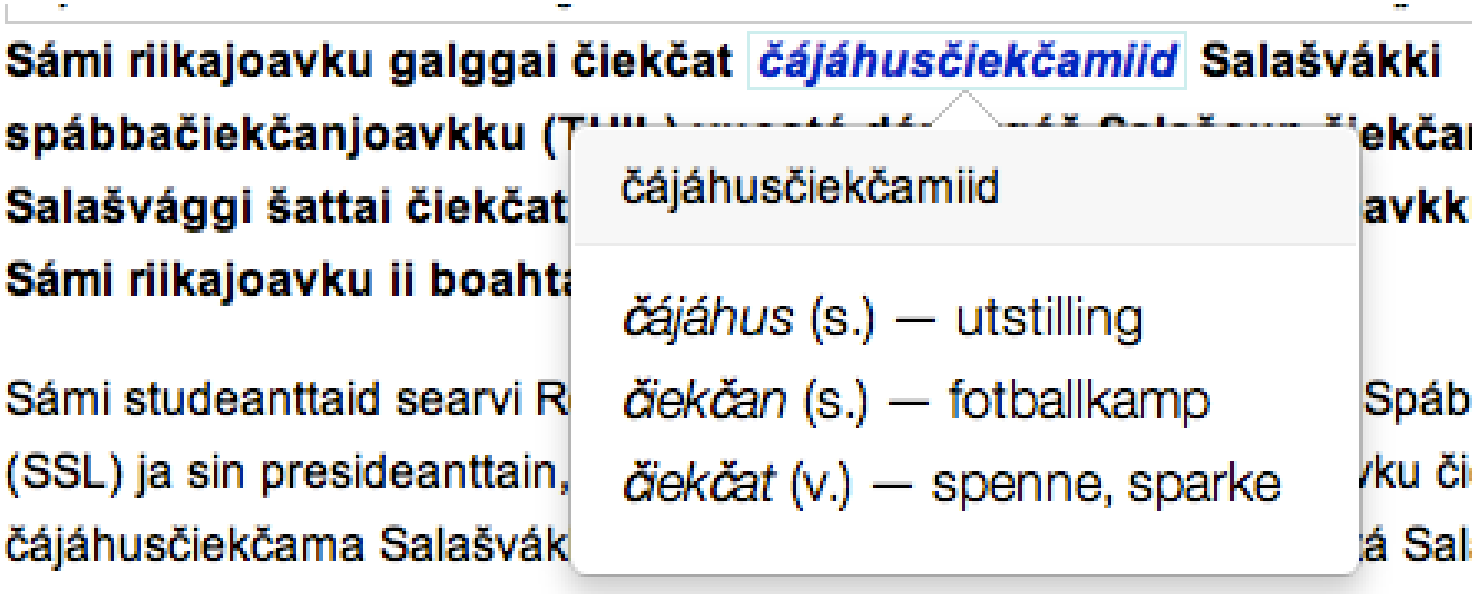
- But for Saami you cannot just do that
- The chance of hitting a word in its dictionary form is low (Antonsen et al 2009):
  - North Saami: 7,9 %
  - Finnish: 10,0 %
  - Norwegian 30,5 %
- So, what is the solution?
- `http://kursa.oahpa.no/2014/12/14/ivgogakti/`

Saami word structure is rich and complex

- Inflection:
  - Case: *niidii* → *nieida* girl
  - Person-tense: *eahtsa* → *iehtsedh* love
- Derivation:
  - Passivisation: *juhkkojuvvui* → *juhkat* drink, *juohkit* share
- Compounding:
  - *bargojoavku* → *bargu* + *joavku* work + group
    - A quarter of all North Saami nouns are compounds
- Cliticisation:
  - *bođiige* → *boahtit* + *ge* came + too

The dictionary understands inflected forms

Click in text



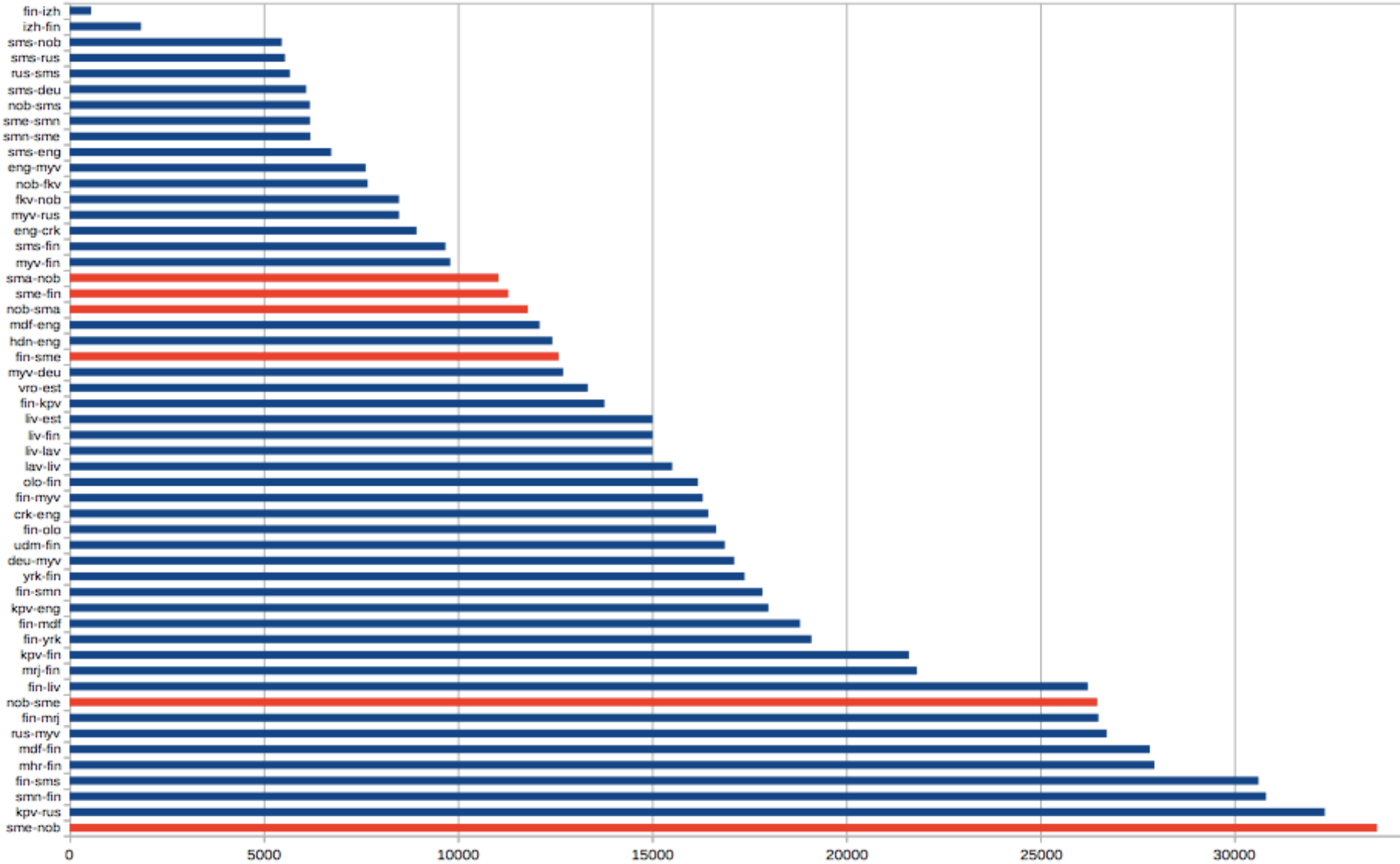
Dynamic FST dictionaries

FST-analysis and translation via a web service

... and may generate word paradigms and complex articles

`http://sanit.oahpa.no/sme/nob`
`http://sanit.oahpa.no/nob/sme`

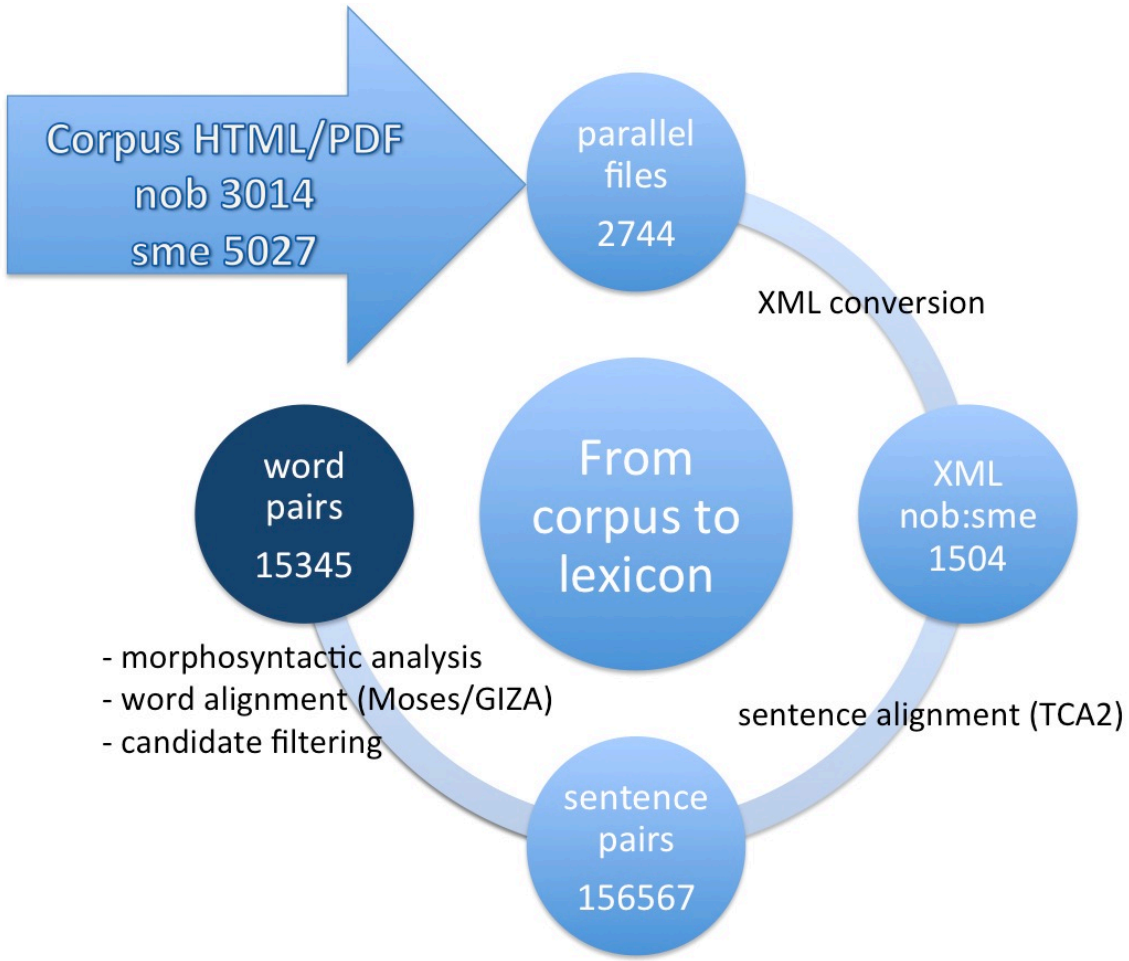

North Haida
Plains Cree
English
Cherokee
Nenets
Kven
Skolt S.
Komi
Udmurt
North S.
Inari S.
Olonets
Meadow Mari
Veps
Hill Mari
Finnish
Ingrian
Erzya
South Saami
Moksha
Votic
Estonian
Võru
Russian
Norwegian
Livonian
Latvian
German


fin-izh
izh-fin
sms-nob
sms-rus
rus-sms
sms-deu
nob-sms
sme-smn
smn-sme
sms-eng
eng-myv
nob-fkv
fkv-nob
myv-rus
eng-crk
sms-fin
myv-fin
sma-nob
sme-fin
nob-sma
mdf-eng
hdn-eng
fin-sme
myv-deu
vro-est
fin-kpv
liv-est
liv-fin
liv-lav
lav-liv
olo-fin
fin-myv
crk-eng
fin-olo
udm-fin
deu-myv
yrk-fin
fin-smn
kpv-eng
fin-mdf
fin-yrk
kpv-fin
mrj-fin
fin-liv
nob-sme
fin-mrj
rus-myv
mdf-fin
mhr-fin
fin-sms
smn-fin
kpv-rus
sme-nob
0
5000
10000
15000
20000
25000
30000

Sources

- Dictionary sources
  - NS ↔ No Jernsletten: Álgosátnegirji
  - NS ↔ Fi Álgu etymological database, Hki
  - SS ↔ No Brunstad and Jåma
  - Other dictionaries A wide array of dictionaries
- Corpus sources
  - Parallel corpora The FAD project

Building dictionaries from parallel corpora (the FAD project)



What can the dictionary logs tell us?

galaga False SoMe nob 2016-02-28T22:54:37 91.100.98.168

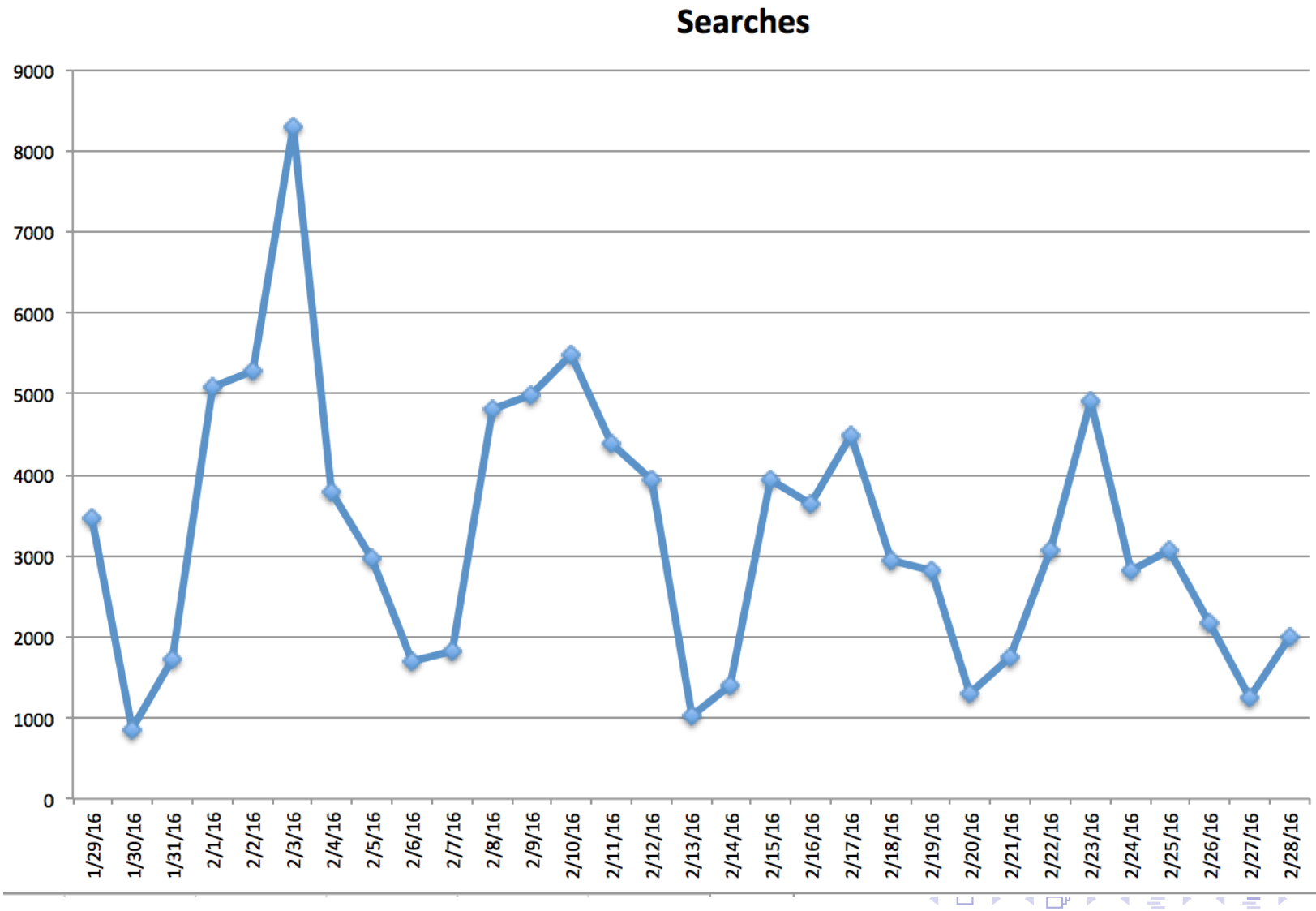
cizážat True cizáš småfugl sme nob 2016-02-28T23:02:01 158.39.204.65

http://gtweb.uit.no/all_nds.html

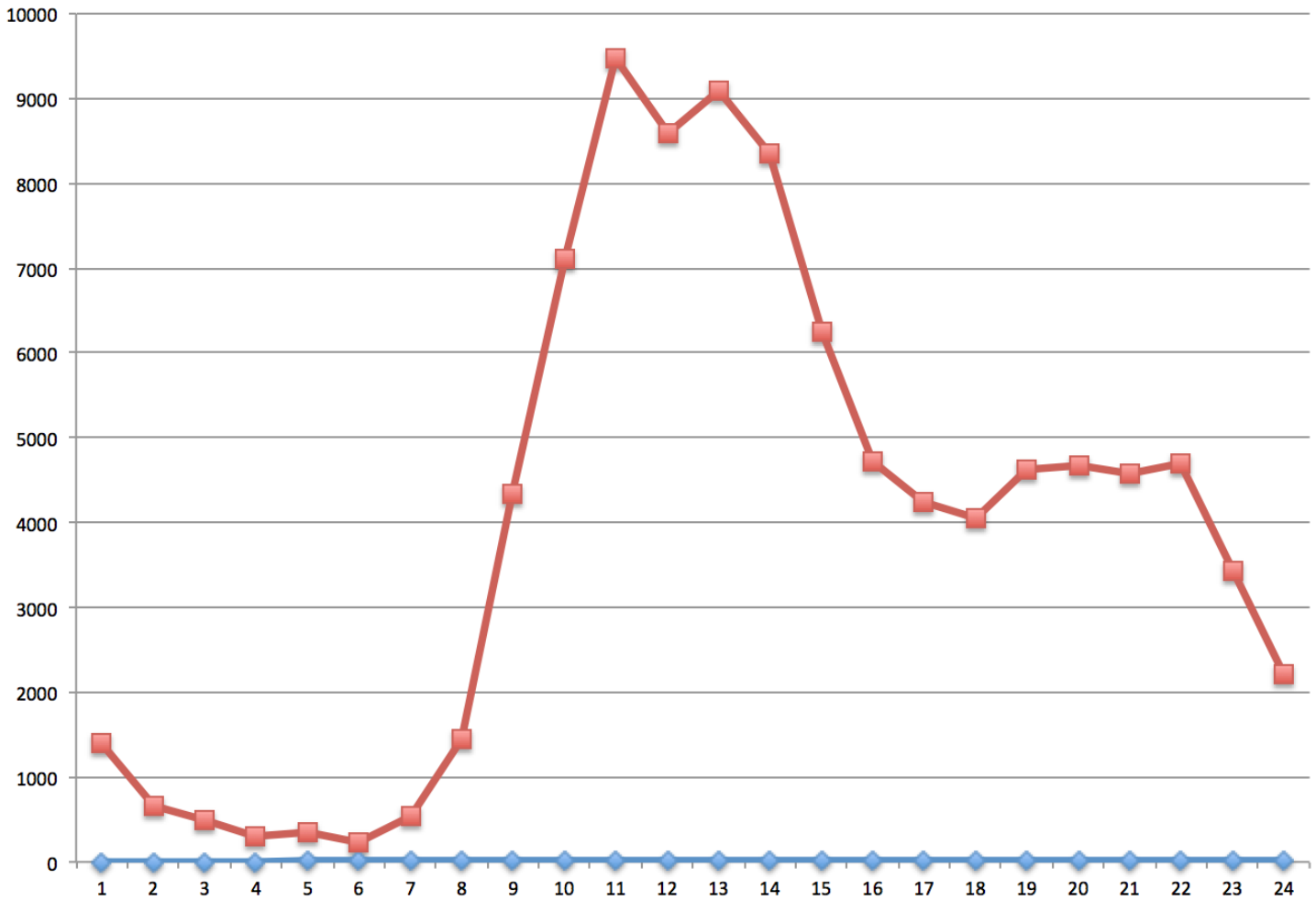
NS - Fi		NS - No		Fi - NS		No - NS	
1224	Jyväskylä	9653	Alta	1022	Oulu	6898	Alta
704	Oulu	3674	Tromsø	561	Rovaniemi	2539	Tromsø
587	Rovaniemi	1644	Tomasjorda	517	Helsinki	1250	Tomasjorda
488	Helsinki	1547	Kautokeino	319	Jyväskylä	982	Kirkenes
284	Varjakka	1082	Birtavarre	120	Haukipudas	820	Oslo
275	Haparanda	957	Oslo	102	Varjakka	681	Kautokeino
257	Muhos	898	Kirkenes	98	Muhos	539	Lakselv
149	Hämeenlinna	882	Varangerbotn	80	Espoo	416	Varangerbotn
139	Sørkjosen	485	Örnsköldsvik	61	Hämeenlinna	323	Birtavarre
122	Kautokeino	406	Stavanger	51	Stockholm	307	Karasjok
99	Sodankylä	392	Kiruna	50	Jokela	256	Lødingen
95	Hauho	308	Trondheim	44	Hyvinkää	214	Älvsjö
85	Espoo	280	Talvik	43	Birtavarre	156	Olderdalen
72	Jokela	259	Luleå	38	Tromsø	150	Storslett
70	Haukipudas	225	Älvsbyn	33	Vantaa	137	Umeå
61	Tromsø	221	Olderdalen	33	Sørkjosen	129	Stavanger
59	Muonio	217	Tromsdalen	32	Tampere	106	Hammerfest
42	Tampere	190	Lakselv	30	Kautokeino	103	Sørkjosen
30	Kalix	189	Umeå	28	Säynätsalo	97	Stockholm
27	Hyvinkää	188	Sørkjosen	28	Imatra	93	Tromsdalen
25	Vantaa	176	Hammerfest	25	Kerava	92	Luleå
23	Oslo	165	Storslett	19	Luleå	82	Arjeplog
19	Alta	155	Bergen	19	Hauho	81	Båtsfjord
18	Stockholm	145	Båtsfjord	18	OEvertornea	78	Storjord
16	Övertornea	134	Karasjok	15	Bergen	66	Narvik
14	Nurmes	117	Hattfjelldal	13	Turku	57	Trondheim
13	Birtavarre	99	Eskilstuna	13	Kempele	52	Hvalstad
12	Kerava	97	Arjeplog	12	Kuopio	46	Borkenes
9	Imatra	90	Stockholm	10	Suomusjärvi	43	Mehamn
9	Bergen	80	Moss	10	Sahalahti	39	Torsken
8	Kittilä	79	Mehamn	9	Oslo	35	Bergen
7	Suomusjärvi	76	Jokkmokk	9	Kolari	33	Rypefjord

SS - No		No - SS	
684	Trondheim	440	Trondheim
525	**Vilhelmina**	226	Hattfjelldal
269	**Umeå**	178	**Vilhelmina**
155	Vestnes	175	**Umeå**
150	Oslo	113	Oslo
119	**Borlänge**	98	Roros
107	**Tomelilla**	89	Molde
99	Hattfjelldal	83	Steinkjer
92	Skreia	83	Grong
85	Singsas	83	**Borlänge**
85	Røros	78	Singsas
80	Steinkjer	64	**Lycksele**
70	**Lycksele**	58	Tromsø
59	Tromsø	56	**Tomelilla**
58	Molde	35	Glåmos
50	**Malå**	34	**Drevsjö**
49	Grong	31	**Östersund**
47	Kolbotn	18	**Arvidsjaur**
45	Stavanger	17	Alta
45	Glåmos	15	**Uppsala**
42	Levanger	13	Lødingen
42	**Drevsjö**	12	**Kiruna**
33	**Östersund**	11	Skreia
27	**Älvsbyn**	11	Levanger
21	Stjørdal	10	Korgen
20	**Uppsala**	9	Stavanger
20	Åsen	9	Horten
16	Mosjøen	8	Østby
15	Røyken	8	Olderdalen
11	Arvidsjaur	8	Lundamo

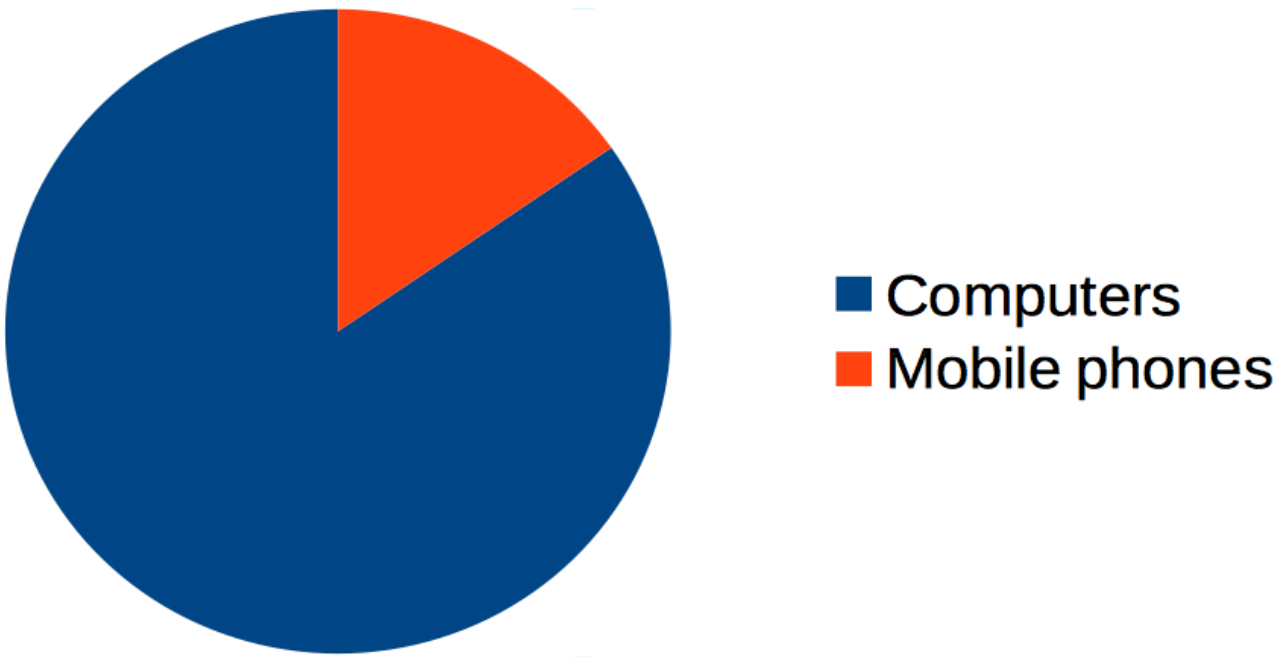
Dictionary use in February: From day to day



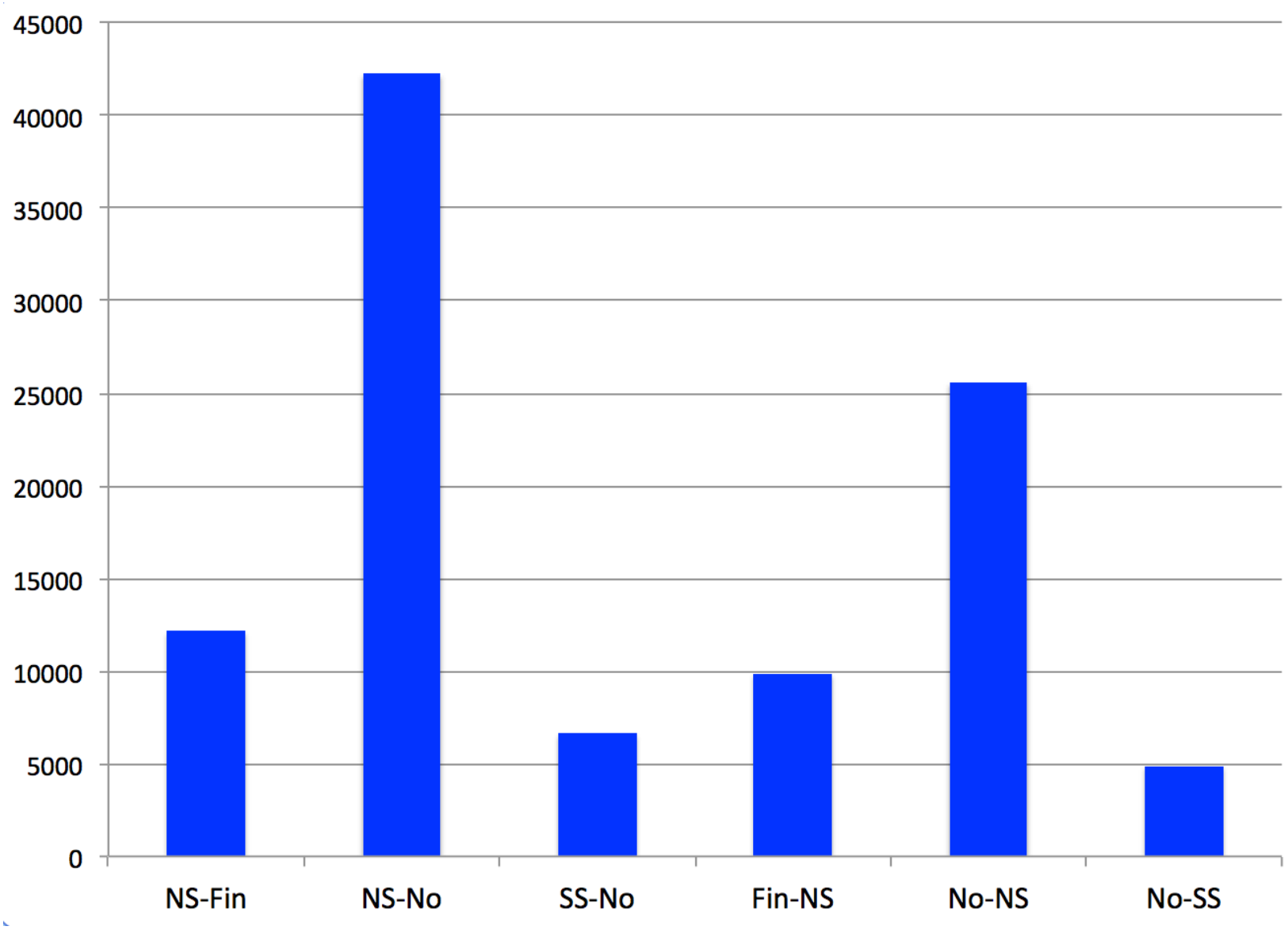
Dictionary use in February: Time of the day



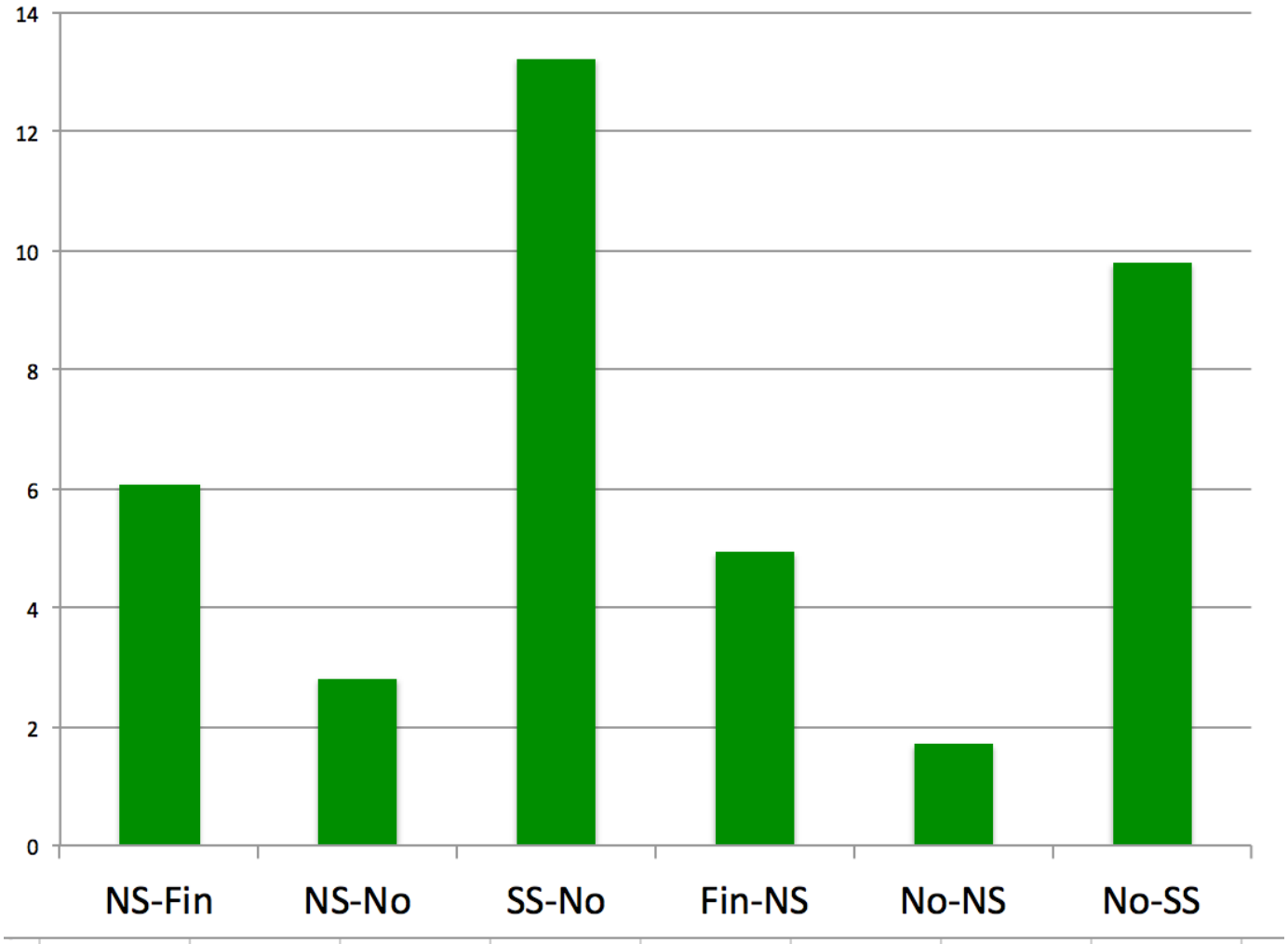
Dictionary use on computers or mobile phones ??



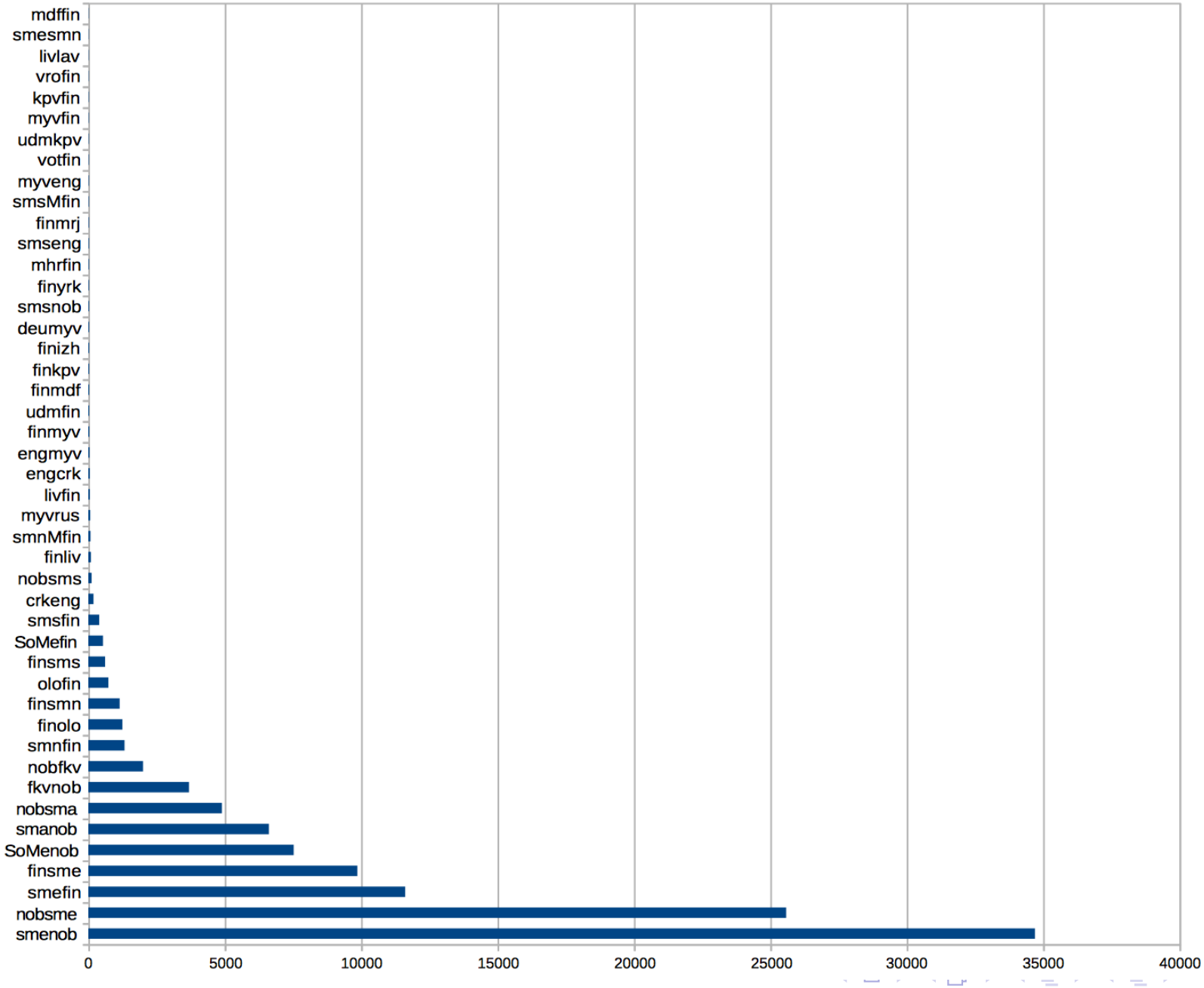
Usage statistics for February



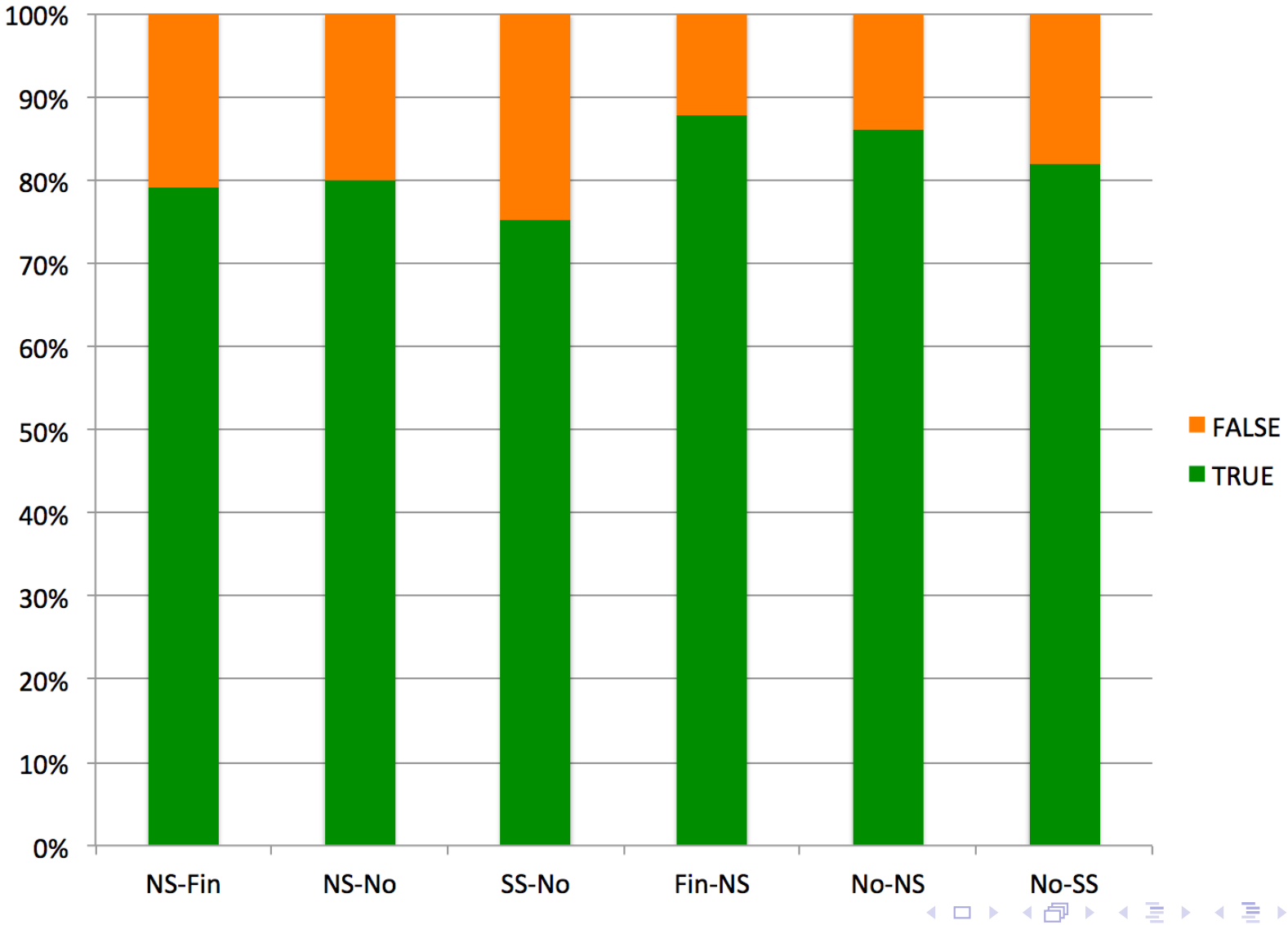
Relative use



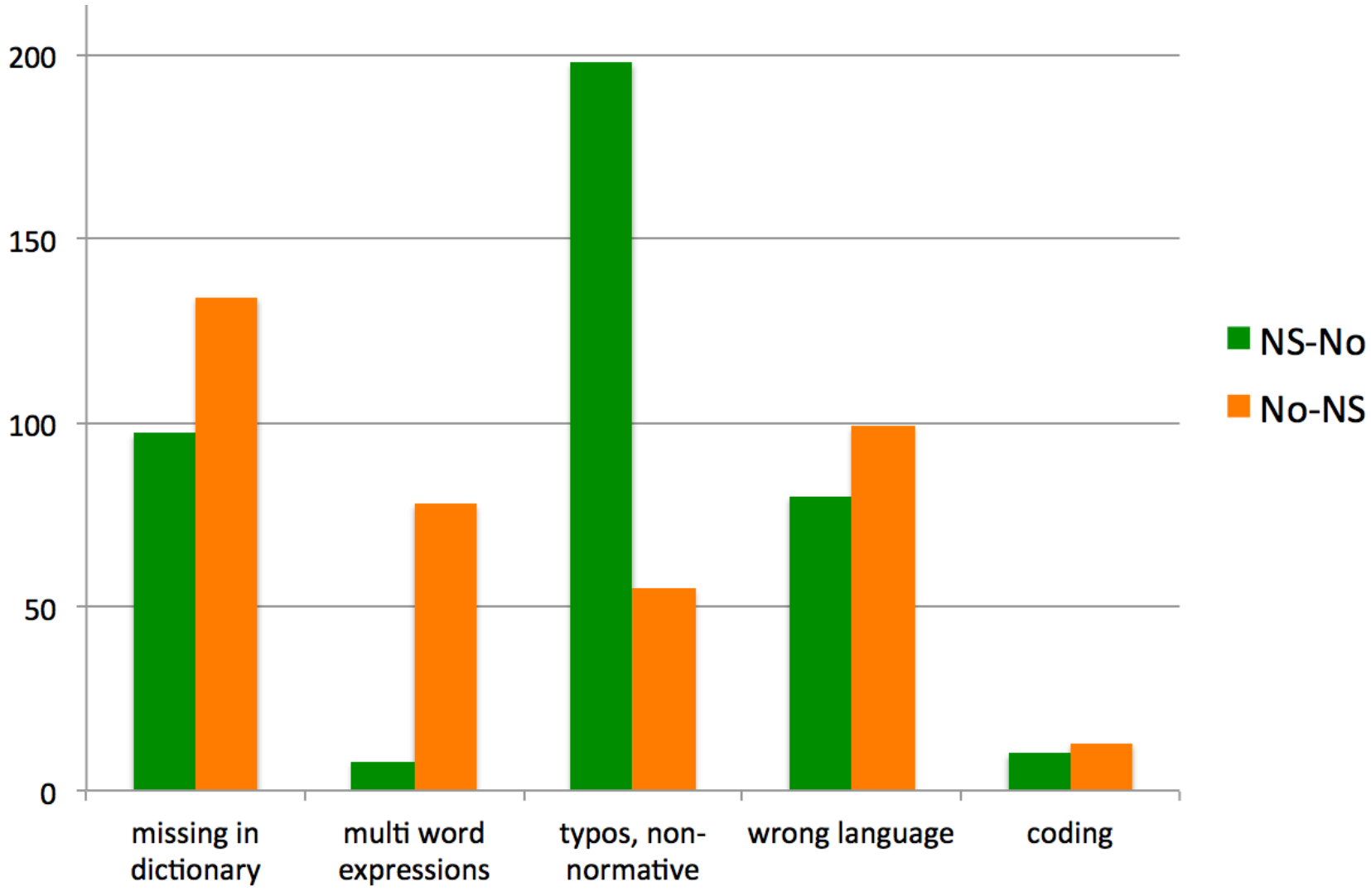
Speakers:
No-NS:
18.000
Fin-NS:
2.000
SS-No:
500


mdffin
smesmn
livlav
vrofin
kpvfin
myvfin
udmkpv
votfin
myveng
smsMfin
finmrj
smseng
mhrfin
finyrk
smsnob
deumyv
finizh
finkpv
finmdf
udmfin
finmyv
engmyv
engcrk
livfin
myvrus
smnMfin
finliv
nobsms
crkeng
smsfin
SoMefin
finsms
olofin
finsmn
finolo
smnfin
nobfkv
fkvnob
nobsma
smanob
SoMenob
finsme
smefin
nobsme
smenob
0
5000
10000
15000
20000
25000
30000
35000
40000

Did the users find the words they were looking for?



Looking at 400 + 400 random false entries



The No-NS false log

Many words are simply missing lilla, innlevering, dømt, egen

9,6% are NS words leat, ruhtačoaggima, mielas

0,8% are Swedish words eta, färg, førsælning, huvudrätt, ihop, kommun, kött, lila, lätt, mösse, någon, omlopp, pastej, salu, ...

Coding errors, some repeat themselves åpne = *Ã¥*pne

Names without capital letter kautokeino, karasjok, tromsø

Words looked up in NDS compared to textbook corpus

- Most overrepresented in the textbook corpus
  - leat be, ja and, mii we/which, čállit write, son s/he, mun I, go QST/than, ii not, sátni word, lohku number, lohkat read, say, count, Elle , tevdnet draw, riekta right, dat it, unni small, govva picture, gávdnat find, don you, galle how many, sárgut draw, bustávva letter, ivdnet colour, ollu much, birra about, around, rehkenastit calculate, dát that
- Most overrepresented in the dictionary log
  - bargat do, work, orrut be, live, mannat go, oažžut get, beassat let, vuolgit leave, háliidit want, vuodjit drive, áigut intend, giella language, boahtit come, šaddat become, čálli writer, fertet have to, olmmoš person, gillet like, viehkat run, galgat should, doalvut remove, čohkkát sit, sápmelaš Sámi

Paradigm lookup vs. corpus frequency

POS	Corp	Corp-%	Parad.	Parad-%	Parad/Corp
Nouns	5903	0.67	3975	0.47	0.7
Verbs	1575	0.18	3281	0.39	2.2
Adjectives	772	0.09	972	0.12	1.3
Pronouns	70	0.01	119	0.01	1.0
Numerals	457	0.05	27	0.00	0.0
Total	8777		8374		

Dictionary users want to inflect adjectives and verbs

How large is the part of the dictionary never accessed?

North Saami - Norw dict: 31764 / 38404 = **82%** never accessed

Norw - North Saami dict: 24950 / 36412 = **68%** never accessed

Which lemma articles should be improved?

- The most popular word class is verb:
  - harder to inflect
  - harder to use
  - harder to point at

Conclusion

- Logging dictionary usage is a good idea
  - Via the log the developers see what words the users look for
- The central POS is the verbs
- The need for L1 $>$ L2 and L2 $>$ L1 coverage is different
  - All L1 words are looked up, but only selected L2 ones
  - This illustrates that dictionary mirroring is not needed
- A large dictionary is not enough,
  - you need to have a relevant language pair, and be present in the language community

Giitu

`http://dicts.uit.no`