Is GF Translation Competitive?

Aarne Ranta
REMU Workshop
Stenungsund 23-24 April 2015

Techniques for MT (and NLP, AI,...)

1948 (Weaver) statistics, engineering

1966 (ALPAC report) grammars, linguistics

1988 (Candide at IBM) statistics

2011 K. Church, "A Pendulum Swung Too Far"

2015

Return of grammars?

yes:

- hybrid systems
- running out of data

no:

- hopes in deep learning
- lack of proper education

Some criteria for machine translation

Quality:

- publication? browsing? post-editing?

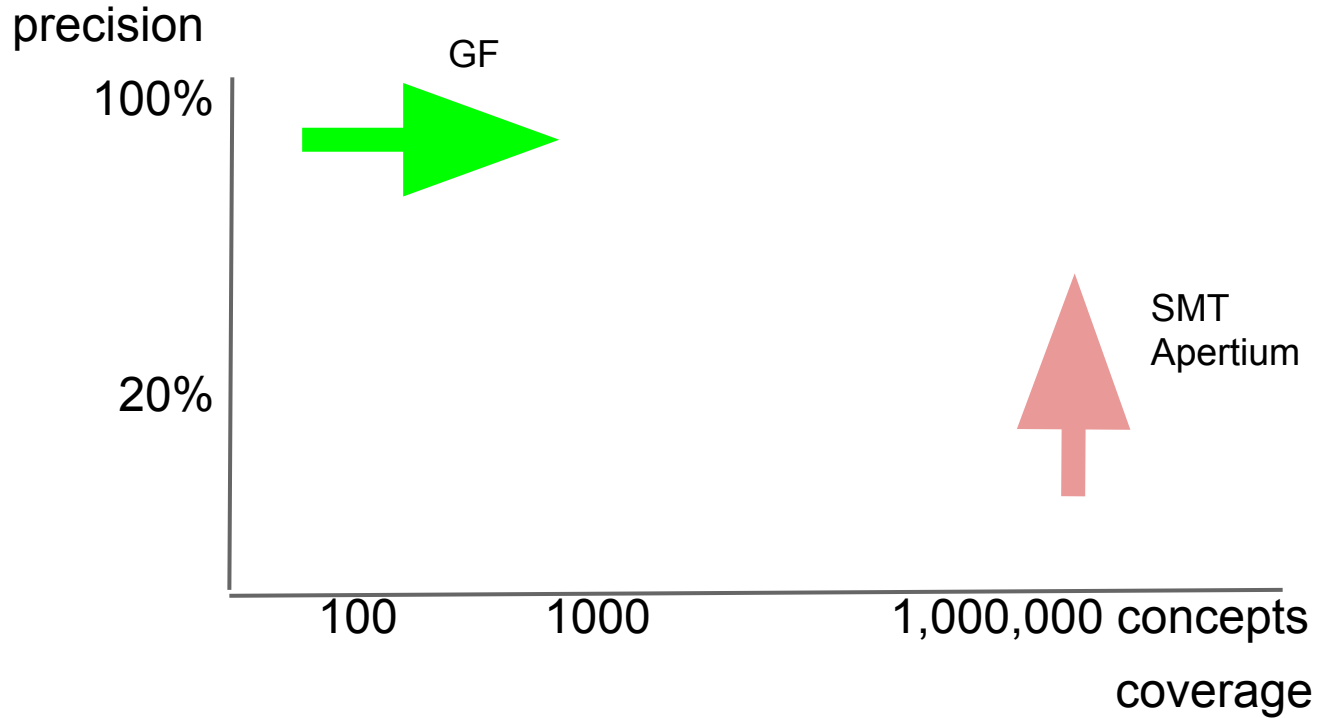Coverage:

- any text? text on a domain? controlled language?

Speed:

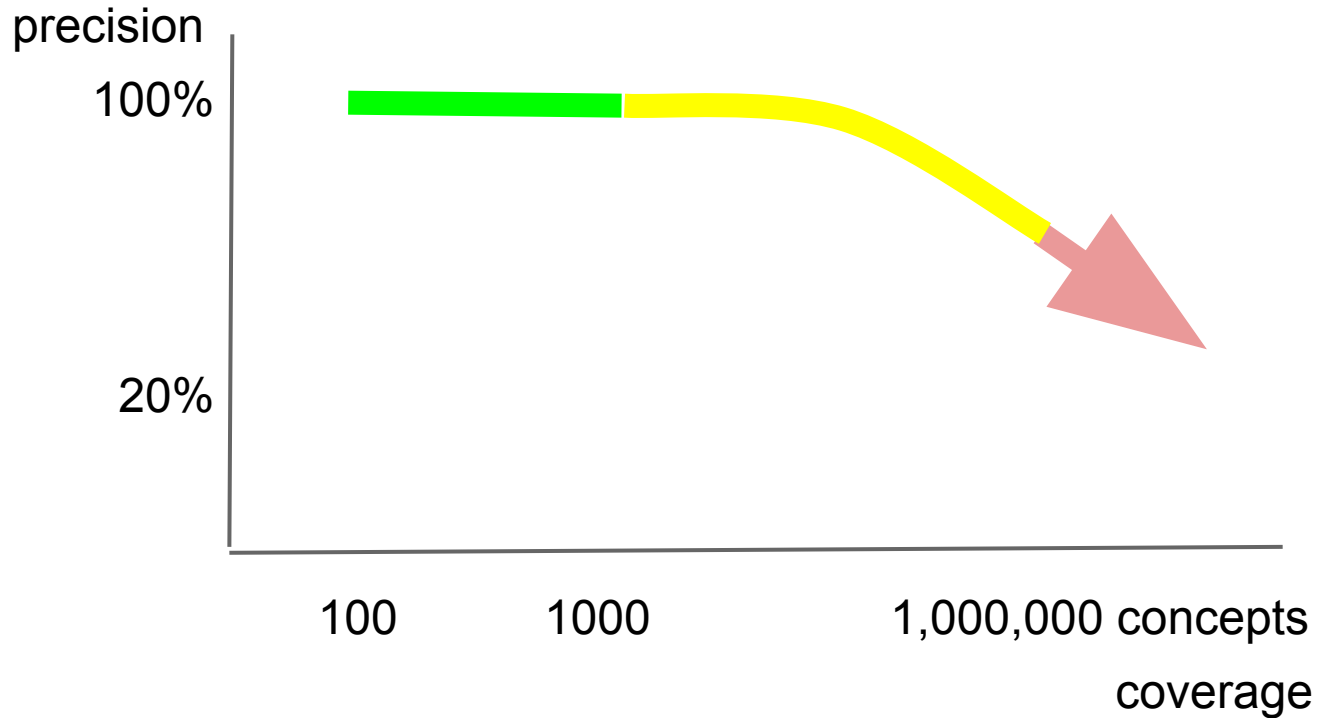- large texts in reasonable time? speech in real time?

Productivity:

- built in days/weeks/months/years/infinity…
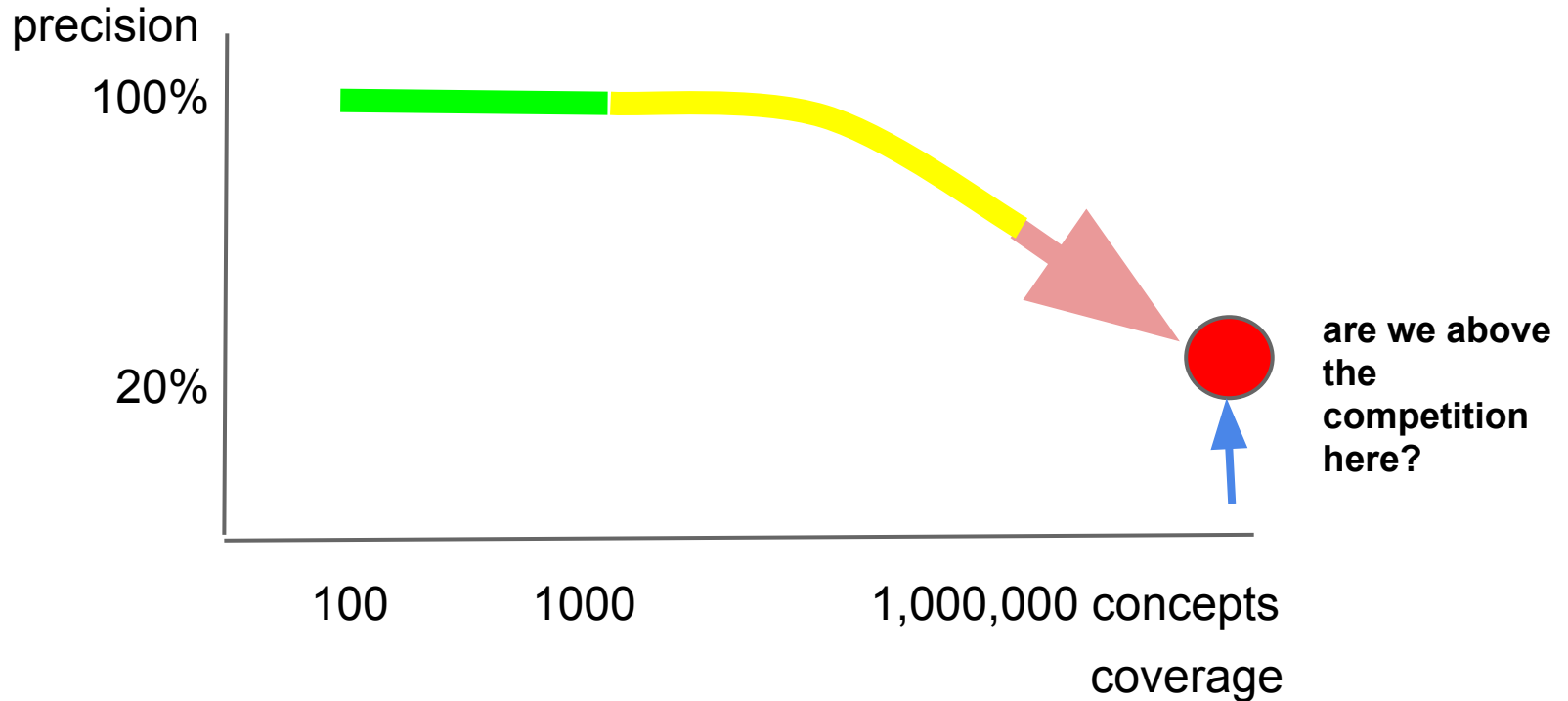- cheaper and quicker than human translation

MOLTO/REMU proposal view

The current GF translator

The current GF translator

Quality evaluation

MT candidate translation *s*

distance *d(s,r)*

manual reference translation *r*

Distance: edit distance, <u>BLEU score</u>, etc

Reference:

- static *(s,r)* pairs from SMT training material
- <u>post-editing</u> *s* to *r* with minimal corrections

Looking at the criteria: quality

Close to publication in CNL (MOLTO use cases)

- post-edit BLEU up to 0.90

Terrible browsing quality worst case

- post-edit BLEU between 0.12 and 0.40
- on the average worse than top-SMT

Post-editing in commercially interesting cases

- post-edit BLEU 0.73 in Tillgänglighetsdatabasen
- possible to post-edit without seeing the source

Some quality figures

<BLEU scores table>

Looking at the criteria: coverage

Yes: we can "translate everything"!

No statistics needed for this: just

- big dictionaries (20 to 70 k lemmas)
- relaxed parsing rules (chunking)
- treatment of unknown words
  - named entity recognition: *Piccadilly Circus*
  - last resort literal chunks: *recieved*

Looking at the criteria: speed

OK for short sentences Eng, Swe, Chi,...

- under 1 second for 10 words

Slow for long sentences and for Fre, Fin, Ita...

<diagram>

Looking at the criteria: productivity

- built in days/weeks/months/years/infinity…
- largely doable without expert knowledge of the languages

The TD case: cheaper and quicker than human translation

- 2 days to build the grammar
- 2 days to post-edit
- human translation at least 6 days
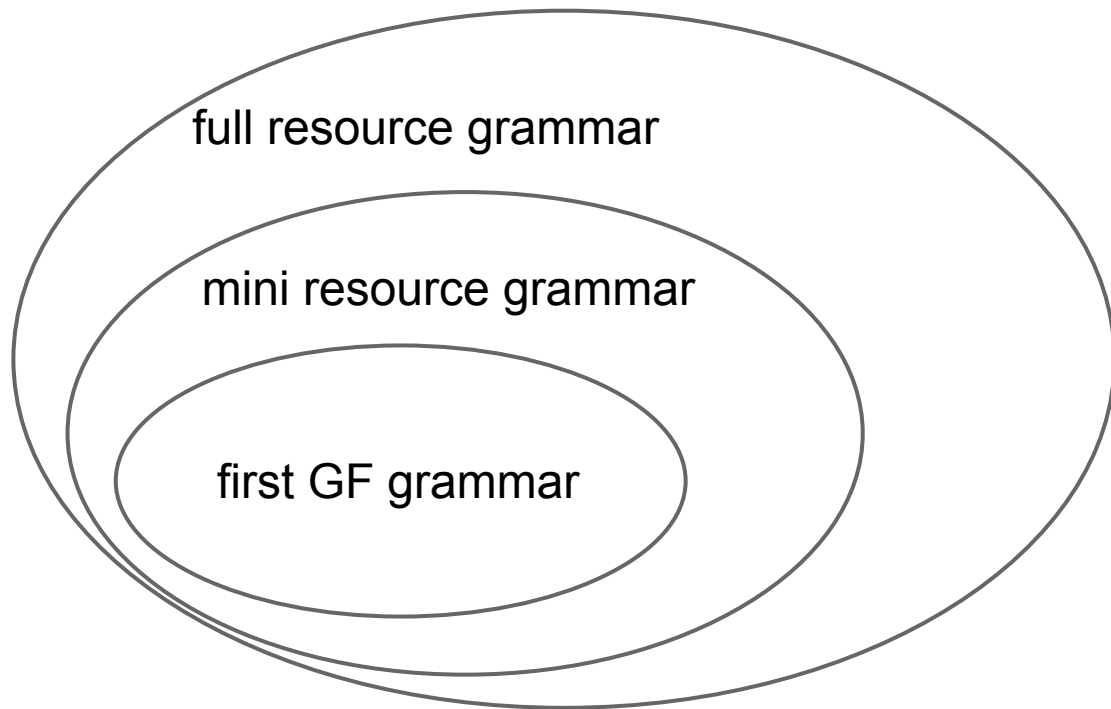- post-editing 2000 words/hour, from scratch 2000 /day

first GF grammar

hours

mini resource grammar
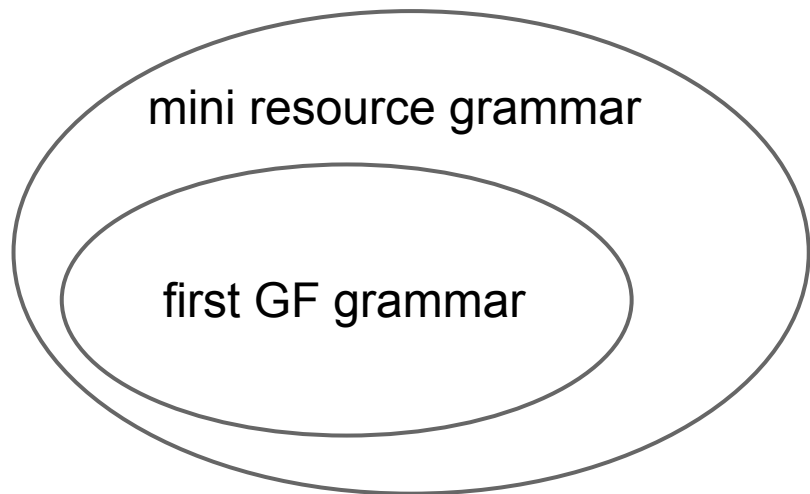
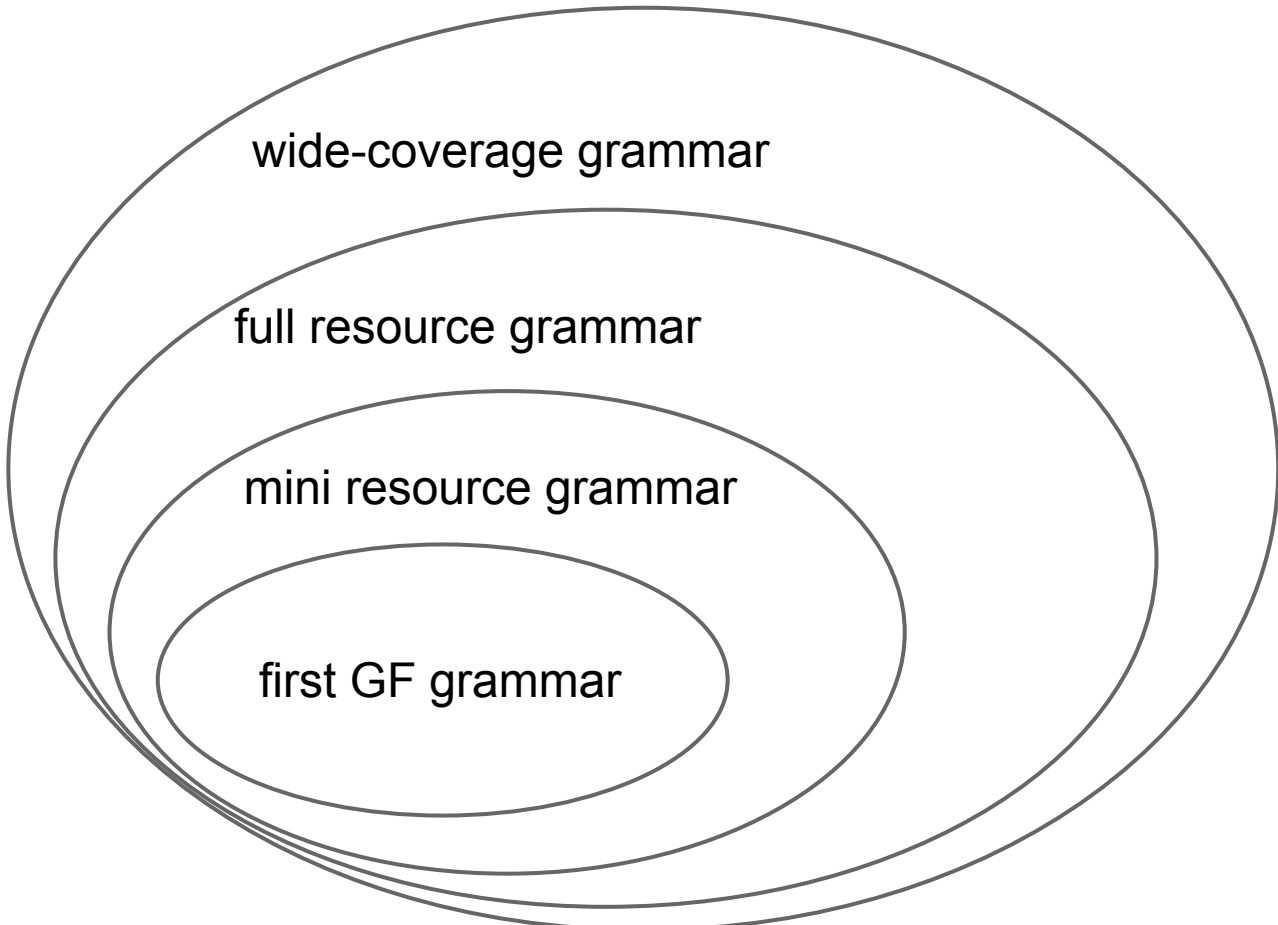first GF grammar

weeks

hours

full resource grammar

mini resource grammar

first GF grammar

months
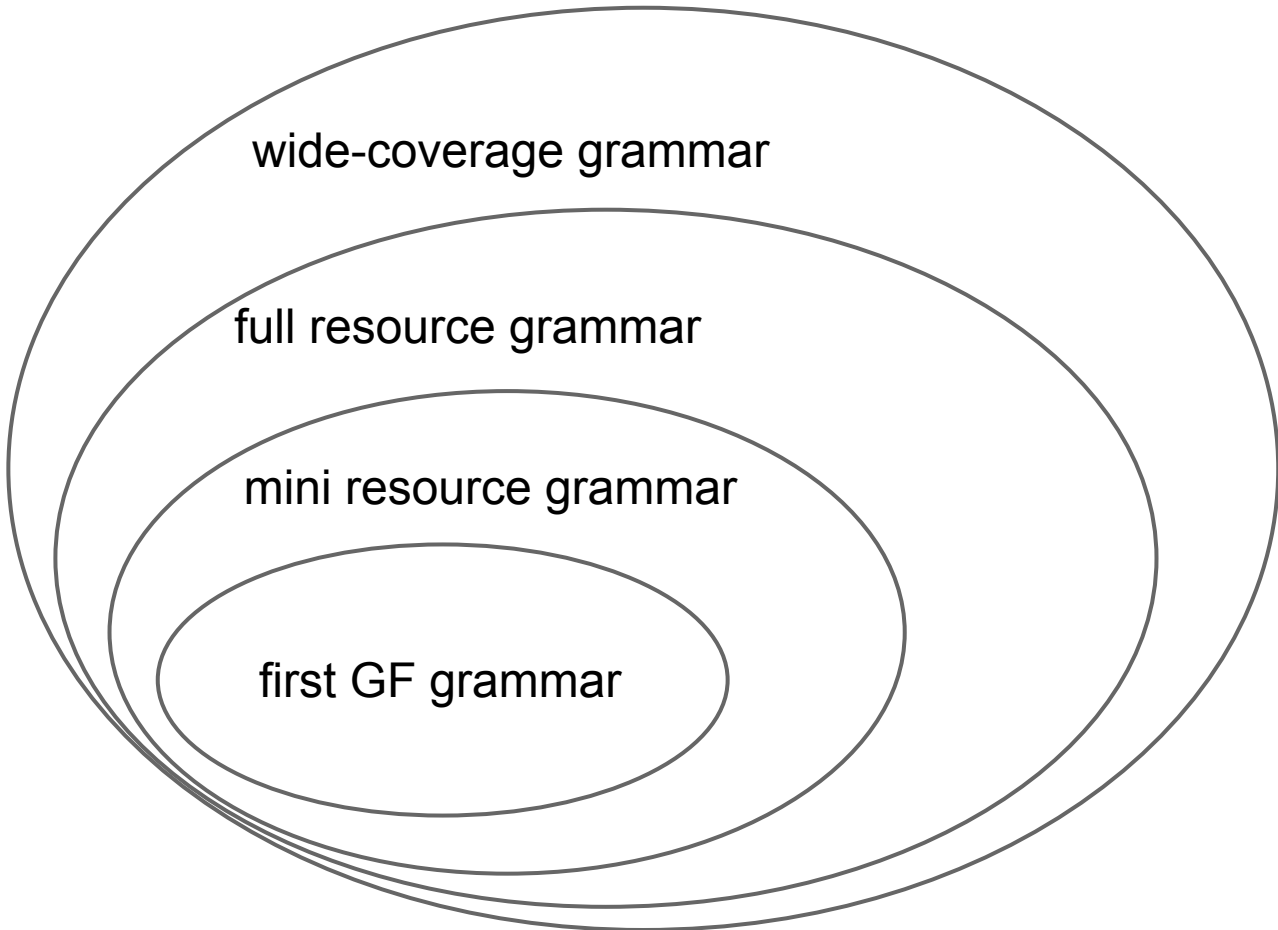
weeks

hours

wide-coverage grammar

full resource grammar

mini resource grammar

first GF grammar

?

months

weeks

hours

wide-coverage grammar

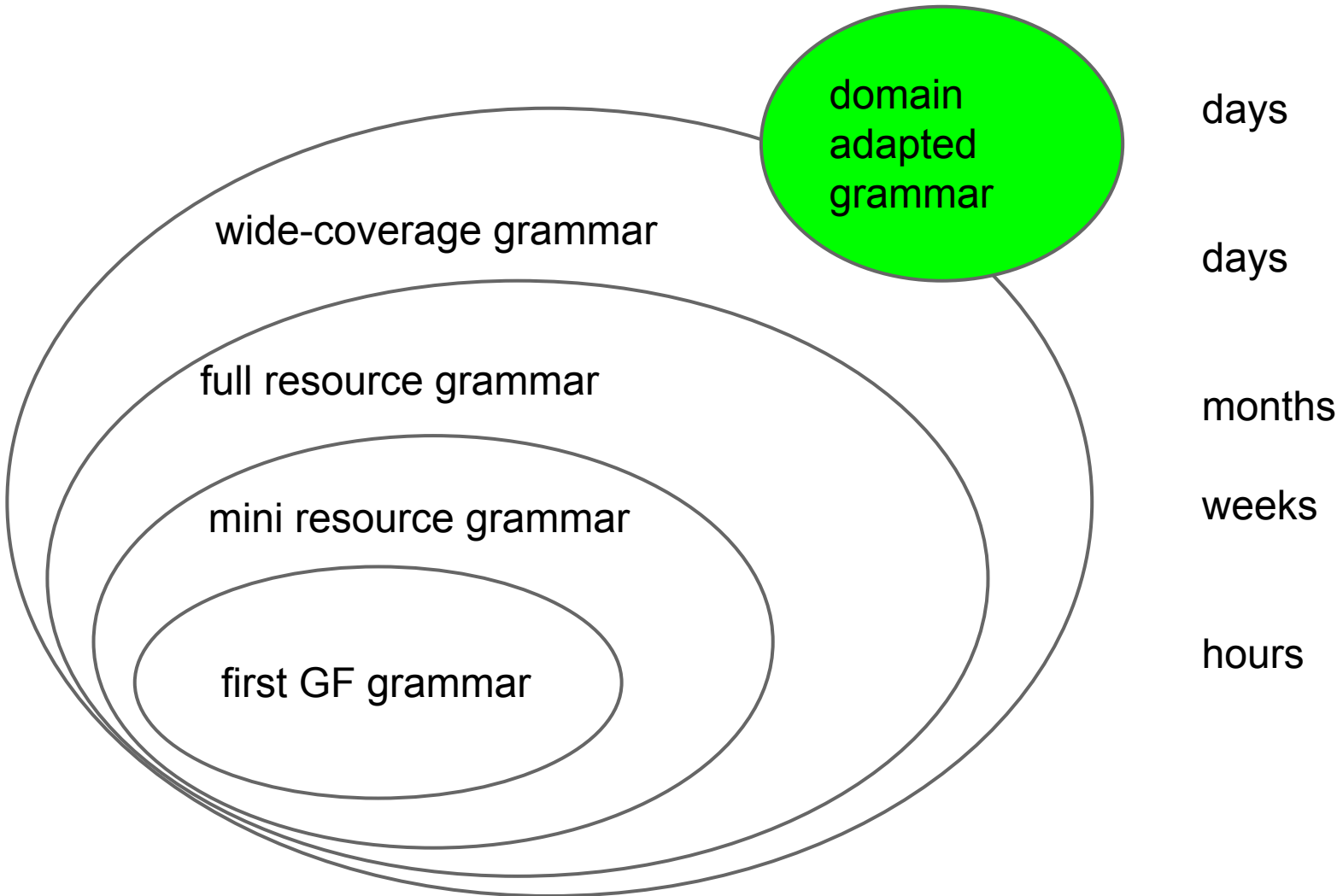**days!**

full resource grammar

months

mini resource grammar

weeks

first GF grammar

hours

domain adapted grammar — days

days

wide-coverage grammar — months

full resource grammar

mini resource grammar — weeks

first GF grammar — hours

Summary of strengths

	GF	SMT
quality	long distance; controlled quality	short distance; random quality
coverage	+	+
speed	-	+
productivity	+ rich morphology, little data	+ idioms, ample data

Can we join forces in a hybrid system?

MT competitions

Common scenario:

- parallel development data
- source language test data
- post-edit BLEU evaluation

Two on-going competitions:
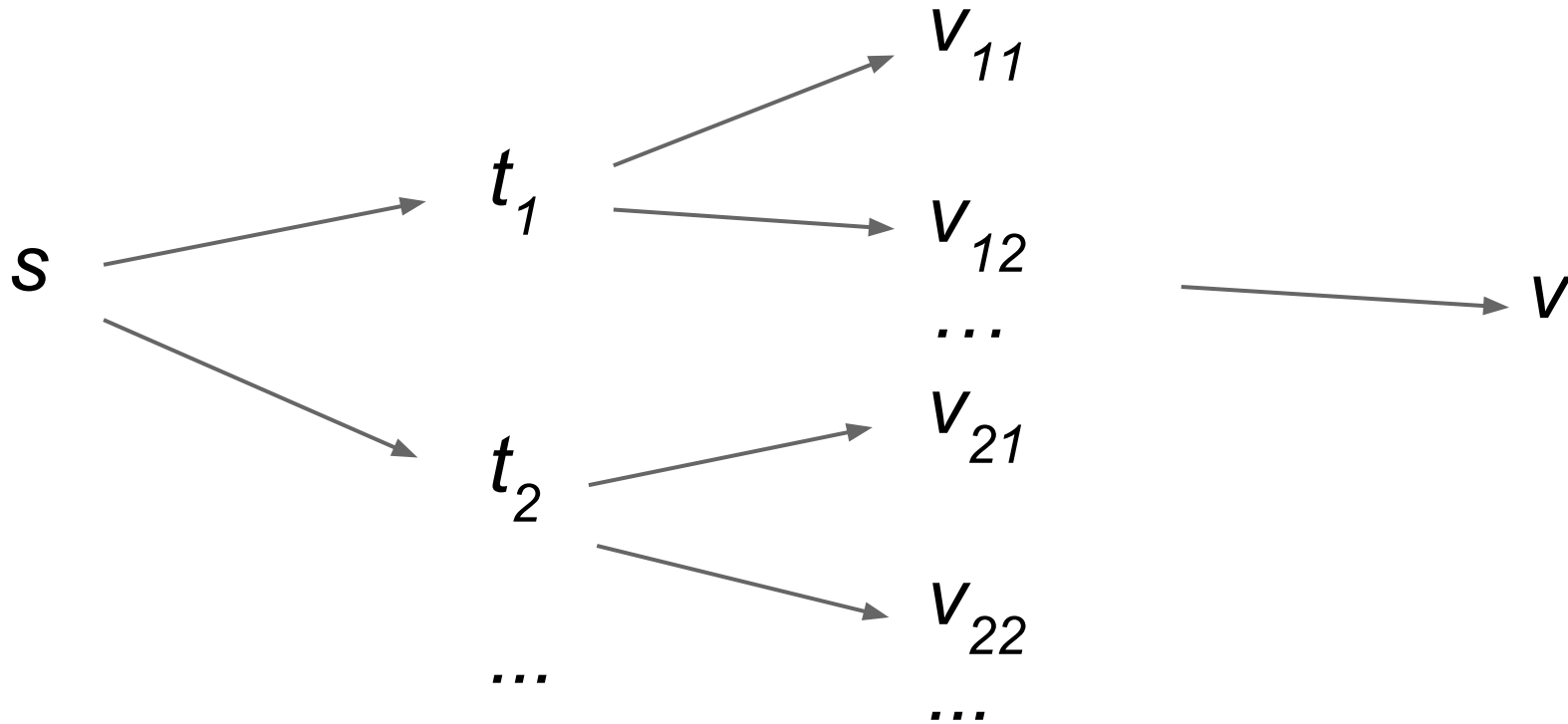
- WMT English-Finnish
- DiscoMT English-French pronoun translation

Idea tested in English-Finnish

parsing   linearization   language model

$$s \rightarrow t_1 \rightarrow v_{11}$$

$$t_1 \rightarrow v_{12}$$

$$\ldots$$

$$s \rightarrow t_2 \rightarrow v_{21}$$

$$t_2 \rightarrow v_{22}$$

$$\ldots \rightarrow v$$

$$\ldots$$

Rationale

Finnish GF grammar is good for generation

- morphology, syntax, lexicon

Probabilistic tree model is insufficient

- context-free
- no distinction between variants

Target language n-gram model helps

- local context for word probabilities

Pronoun translation

1. *I see a tree. **It** is green.*

    *Je vois un arbre. **Il** est vert.*

2. *I see a house. **It** is green.*

    *Je vois une maison. **Elle** est verte.*

The context

Nouns and names in earlier sentences:

… tree … house …

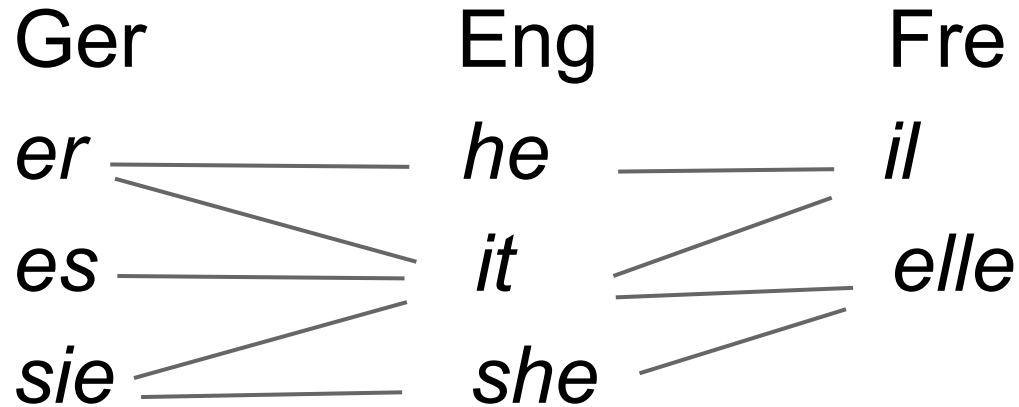Their syntactic context

My house has no door. It only has windows.

The usage context

The monkey ate the banana because
- *it was hungry   er hungrig war*
- *it was ripe   sie reif war*
- *it was lunchtime   es Mittag war*

The general picture

Ger   Eng   Fre

er ——————— *he* ——————— *il*

es ——————— *it* ——————— *elle*

sie ——————— *she*

The reference decides!

Gender-dependent pronoun

```
fun Pron : N -> Pron

lin Pron n = case n.g of {
    Masc => "il" ;
    Fem => "elle"
    }
```

Our algorithm

1. Store the abstract syntax trees of *n* previous sentences.
2. List possible pronoun referents in these trees.
3. Replace each pronoun with `Pron` *R* where *R* is the most likely referent.
4. Generate target language from this tree.

Pronoun translation

1. *I see a* ==*tree*==. ***It*** *is green.*

   Pron tree_N

   Je vois un arbre. ***Il*** *est vert.*

2. *I see a* ==house==. ***It*** *is green.*

   Pron house_N

   Je vois une maison. ***Elle*** *est verte.*

What we expect

We don't expect to win (yet)!

But this is a new kind of a task for GF, and we expect to
- get new ideas
- make GF translation better known