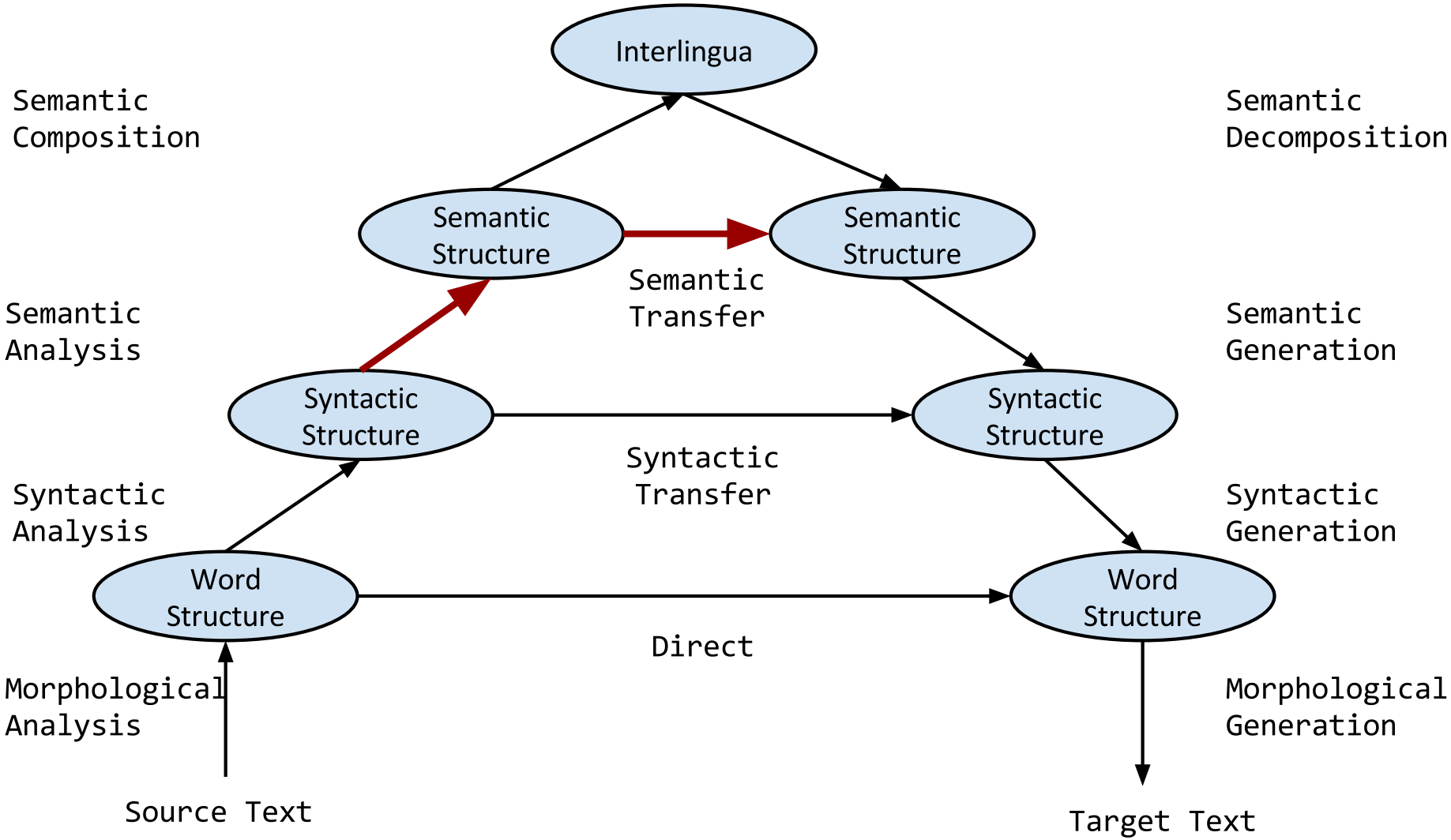
Lexical-Semantic Features in Cross-Lingual Applications

Yulia Tsvetkov

Joint work with Chris Dyer

Language Technologies Institute
Carnegie Mellon University

The Vauquois Triangle for MT

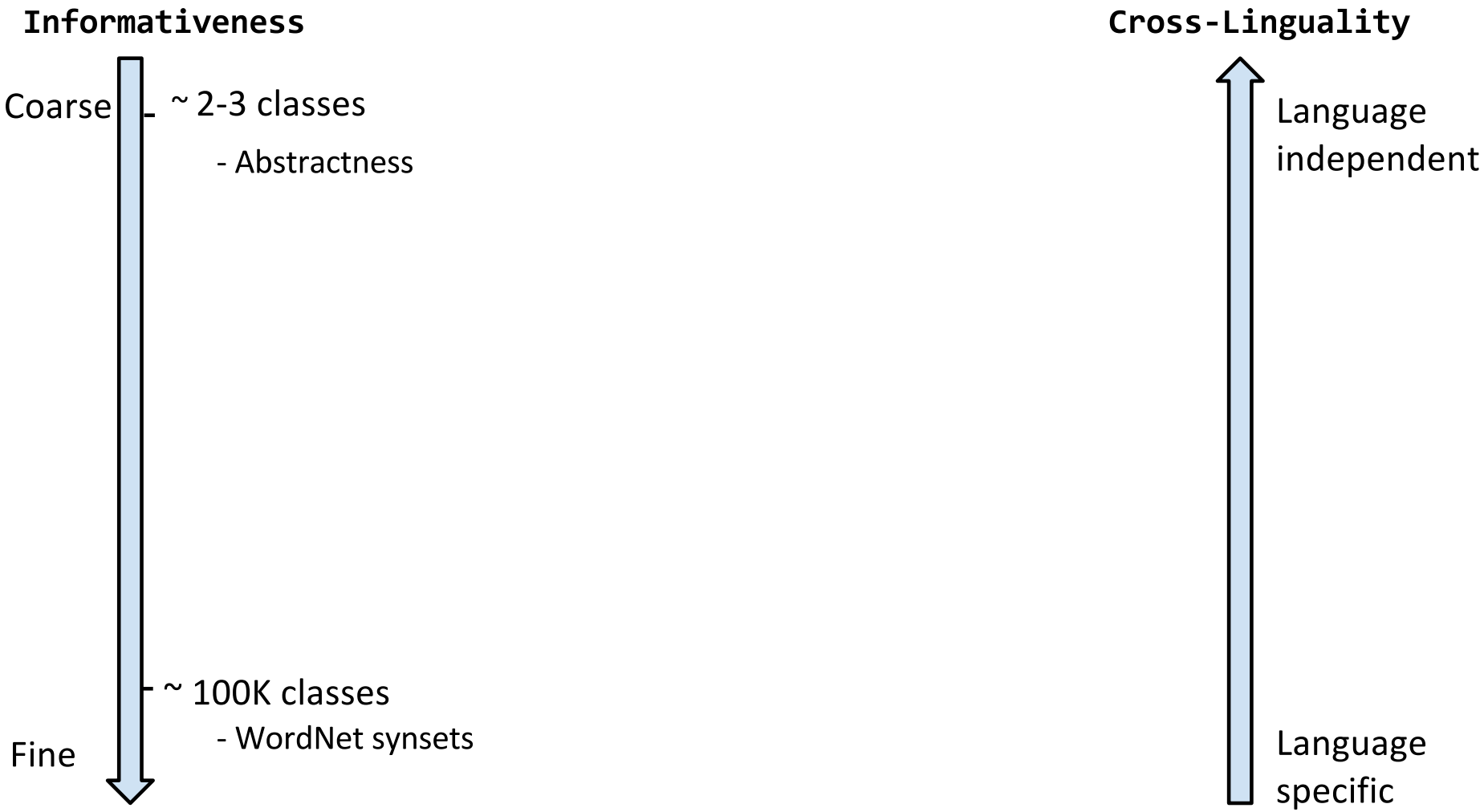


(Dorr et al. 2004)

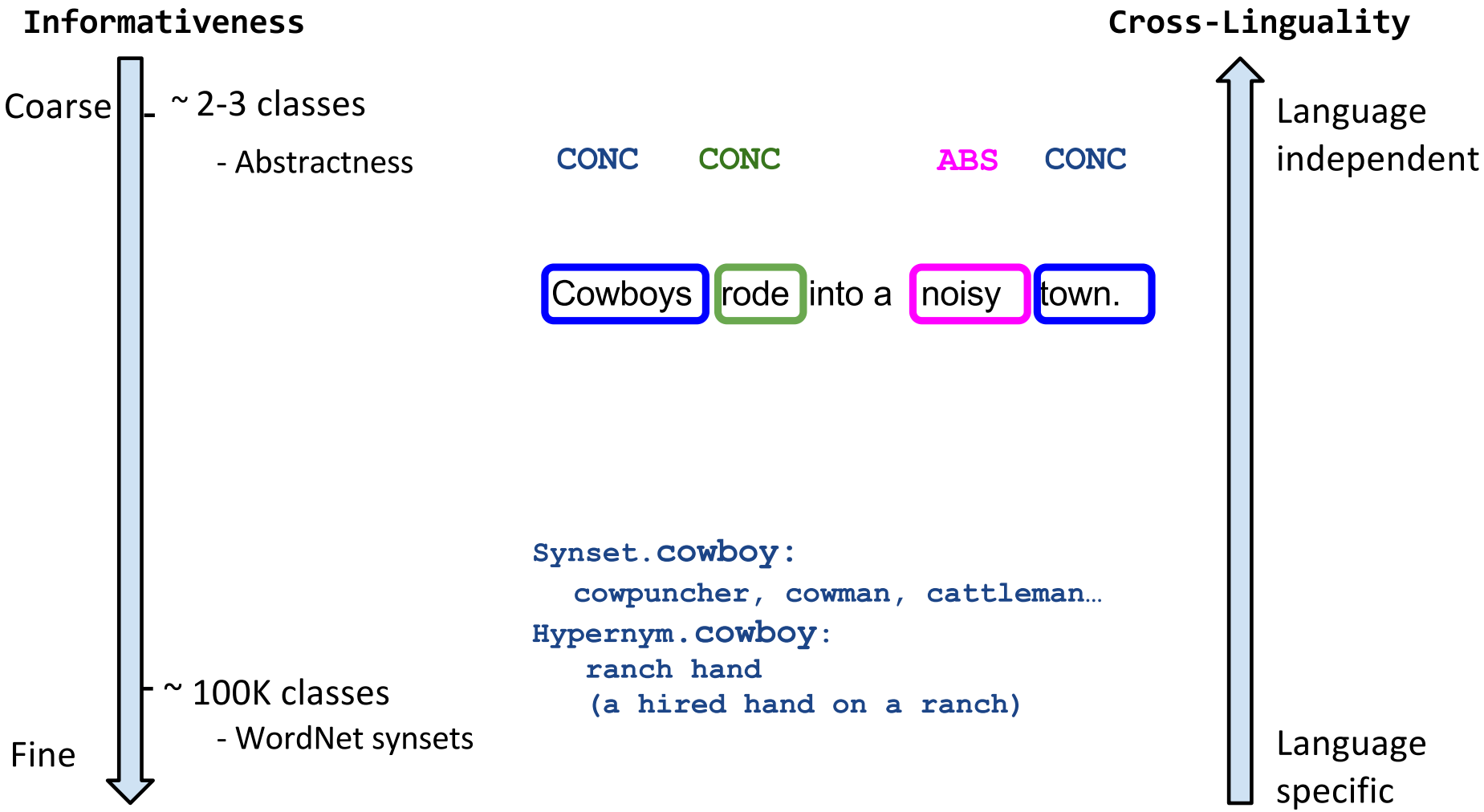
Outline

1. Lexical-semantic features
2. Construction of resources
3. Cross-lingual applications

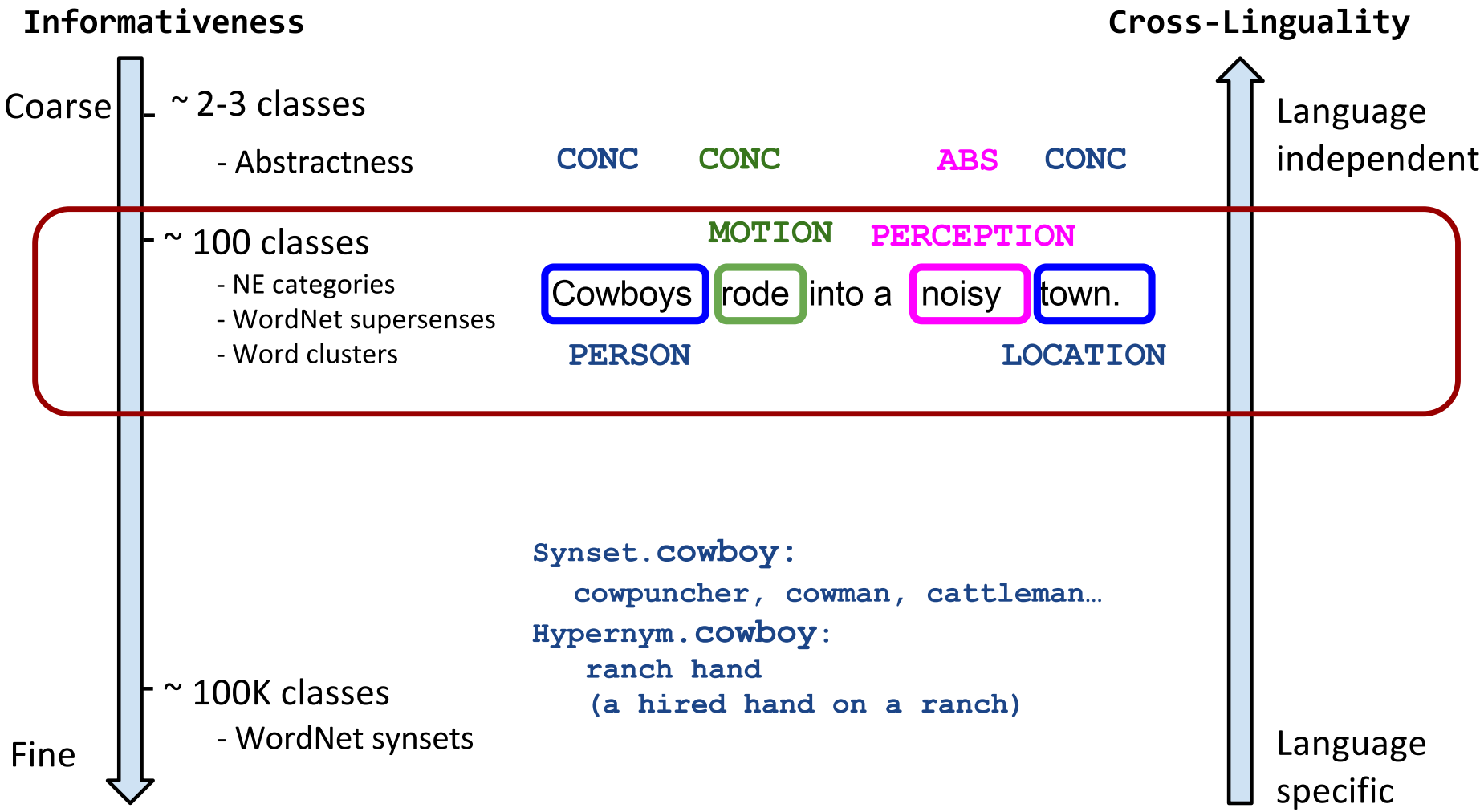
Lexical-semantic features



Lexical-semantic features



Lexical-semantic features



WordNet supersenses

- A coarse form of **word sense disambiguation**; generalizes NER beyond proper names
- Categories broadly applicable across domains (Schneider et al. 2012), and across languages (Schneider et al. 2013, Tsvetkov et al. 2013,)
- WordNet (Fellbaum 1998) has 26 classes for nouns, and 15 classes for verbs

Supersenses	Example words
ARTIFACT	bridge, restaurant, toaster, aspirin
LOCATION	downtown, stage, left, India, airspace
FOOD	juice, apple
COMMUNICATION	discussion, contract, proposal
COGNITION	puzzlement, intuition, awareness
MOTION	fly, settle, lift
EMOTION	love, envy
CONTACT	lean, attach, knock

Adjective supersenses

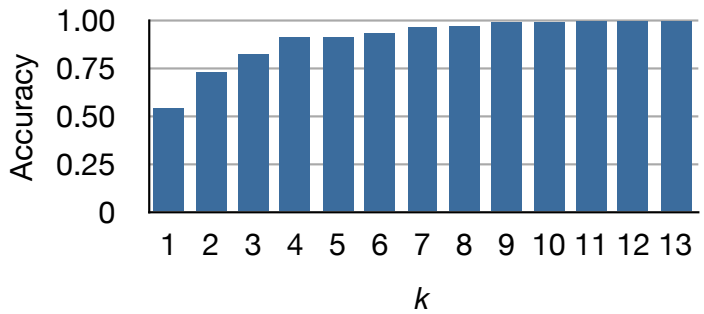
- Unlike nouns and verbs, there is no hierarchical taxonomy of adjectives in English WordNet
- German WordNet — GermaNet (Hamp and Feldweg, 1997) has such a taxonomy with 13 classes for adjectives, but its content is proprietary

Words	Supersenses	Description
purple, shiny, noisy	PERCEPTION	color, lightness, taste, smell, sound
compact, gigantic, far	SPATIAL	dimension, direction, origin, shape
creamy, metallic, dry	SUBSTANCE	consistency, material temperature
bossy, adept, popular	BEHAVIOR	character, inclination, discipline, skill
cheesed off, cheerful	FEELING	feeling, stimulus
chaotic, similar, vague	MISC.	order, completeness, validity

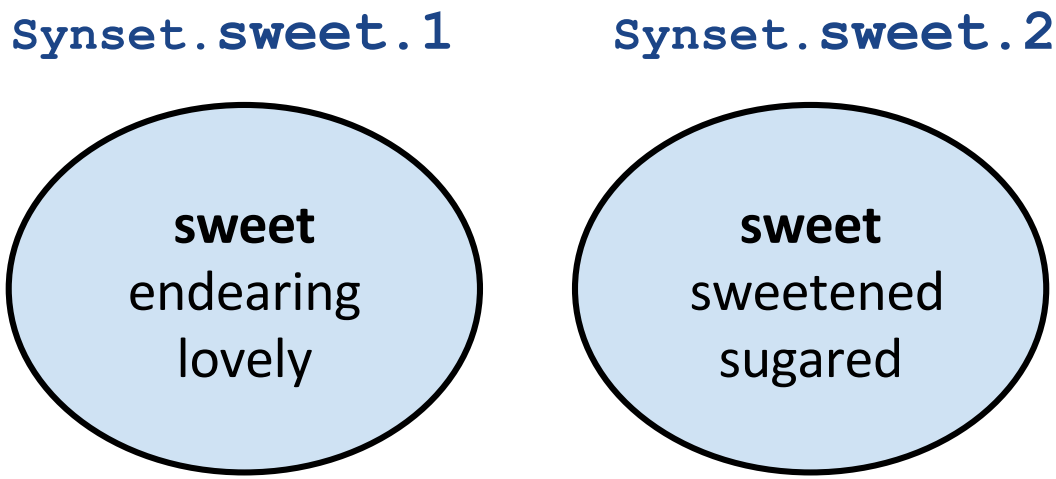
Adjective supersense classifier

We train a weekly supervised multi-class classifier that labels adjective *types*:

- **Labels** – 13 coarse semantic classes from the GermaNet taxonomy.
- **Training data** – Manually annotated 1K English adjectives.
- **Features** – Vector space word representations built from an unlabeled corpus. Vector construction is a multilingual variation of a traditional LSA (Faruqui and Dyer, forthcoming).
- **Model** – Random forest
- **Evaluation** – Accuracy-at-k. For $k = 4$, the classifier accuracy is 91%



Adjective supersenses



- **Released resources** – 10,793 supersense-labeled words (without context); 5,181 supersense-labeled WordNet synsets (in context)

Augmenting English Adjective Senses with Supersenses.
Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, Chris Dyer. Submitted to LREC.

Outlook

- Cross-lingual semantic features
- Construction of resources
- Cross-lingual applications
  - English-Swahili noun supersenses
  - English-Swahili MT
  - Cross-lingual metaphor detection

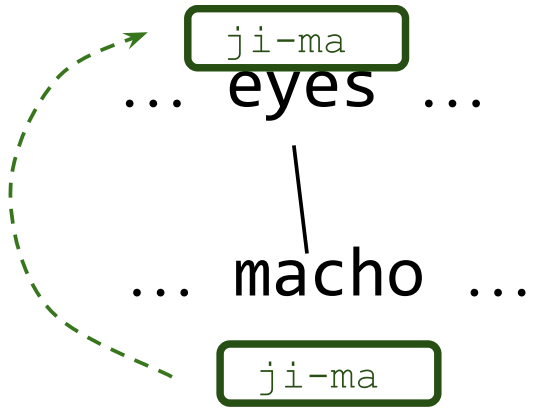
Case Study 1: Swahili noun classes

- Swahili has 9 noun classes which are, similarly to other Bantu languages, semantically-based, and are similar to supersenses.
- Different noun classes have distinct morphology, which affects not only nouns but also their modifiers (adjectives, numbers, demonstratives) and verbs.

Words	Supersenses	Class description
m-tu(person), wa-tu(people)	M-WA	people, animals, insects, birds, fish
m-ti(tree), mi-ti(trees)	M-MI	trees, plants, body parts
ji-cho(eye), ma-cho(eyes)	JI-MA	plural and mass nouns, fruits, round things
ki-tabu(book), vi-tabu(books)	KI-VI	artefacts, names of languages

English-Swahili noun classes

We train a classifier that labels English nouns with Swahili classes:



- **Training data** – 1,524 high-confidence English nouns aligned to annotated Swahili nouns.
- **Features** – Supesenses, VSM, Brown clusters.
- **Evaluation** – Accuracy of 87% in 10-fold cross validation.
- **Outcome** – 17,291 English nouns annotated with Swahili classes.

Case Study 2: English-Swahili SMT

- EN $\rightarrow$ SW, 14K paralles sentences, 4-gram LM
- cdec (Dyer et al., 2010) decoder with MIRA (Chiang, 2012) optimizer

	BLEU
Baseline	16.9
+Supesenses, VSM, Brown clusters	17.2

Case Study 3: Cross-lingual metaphor detection

- **Labels** – L (literal)/ M (metaphor) classes.
- **Training data** – ≈3K annotated Subject-Verb-Object (SVO) and Adjective-Noun (AN) constructions.
- **Features** – English or projected-to-English Supersenses, Abstractness, VSM.
- **Model transfer** – Train metaphor classifier on English SVO and AN constructions, predict for resource-scarce languages: Russian, Spanish, Farsi

	SVO	AN
EN	0.79	0.85
RU	0.84	0.77
ES	0.76	0.72
FA	0.75	0.74

Metaphor Detection in Low-Resource Languages.
Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, Chris Dyer.

Submitted to EACL.

Future work: Cross-lingual annotation of semantic relations

Freebase knowledge base (Google), has 41M non-numeric entities, 19K properties, and 596M assertions.
Example: *place-of-birth*(BarackObama; Honolulu)

- Now:
  is-literal(Empty; Cup)
  is-figurative(Empty; Sound)
  is-figurative(Пустой; Звук)
- Our goal – cross-lingual annotation of Freebase relations:
  place-of-birth(БаракОбама; Гонолулу)

Thank You!