Supersense Tagging Guidelines

What should be tagged?

What counts as a noun?

For the current phase of annotation, we should be strict about only tagging things that (as a whole) serve as **nouns**. Though semantic categories like ATTRIBUTE (*modifiable*), LOCATION (*southwestern*, *underneath*), RELATION (*eleventh*), and TIME (*earlier*) may seem relevant to adjectives, adverbs, prepositions, or other parts of speech, worrying about those would make our lives too complicated.

Special cases:

- **Anaphora** (pronouns, etc.): if the supersense is clear in contextâ€”e.g. it has a clear nominal referent or obviously refers to a specific category (e.g. *someone* referring to a PERSON)â€”that supersense may be applied; leave blank otherwise (e.g. dummy *it*; *others* if too vague).
    - Never tag WH- or relative pronouns like *who* or *which*.
    - Never tag quantifiers in the gray area between determiners, adjectives, and pronouns: *some*, *all*, *much*, *several*, *many*, *most*, *few*, *none*, *each*, *every*, *enough*, *both*, *(n)either*, and generic senses of *one*. (These quantifiers often show up in partitives: *all/some/none of the X*, etc.)
    - For Arabic annotation we are not supersense-tagging ANY anaphora.
- **Verbal nouns/gerunds**
    - In Arabic, we have decided to tag *masdar* instances as nouns.
- **Mentions** of words (e.g., *The word "physics" means...*) should be tagged as COMMUNICATION because they are about the linguistic item.

Determining item boundaries

It is often difficult to determine which words should belong together as a unit (receiving a single supersense tag) vs. tagged separately. Some guidelines:

- Try to treat **proper names** as a unit. (Lack of capitalization makes this especially difficult for Arabic.)
    - Names of titles SHOULD be included if they appear as they might be used in addressing that person:

        President Obama
        United States *President Obama*
        Barack Obama, *president* of the *United States*

    - Honorific prefixes and suffixes should be included: *Dr. Fred Jelinek, Ph.D.*, *King Richard III*
- Other **multiword phrases** can be treated as a unit if they "go together strongly".
    - For example, *lexical semantics* is a standard term in linguistics and should therefore be considered a single unit. Note that *lexical* is not a noun, but it may be included as part of a term that overall functions as a noun.
    - Indications of whether an expression should be treated as a unit might include: conventionality (is it a particularly common way to refer to something?), predictability (if you had to guess how to express something, would you be likely to guess that phrase?), transparency (if you hadn't heard the whole expression before, would its meaning be clear from the individual words?), substitutability (could you replace a word with a similar word to get an equally normal expression meaning the same thing?).
    - Consider: would you want to include the expression as a unit in a dictionary?

Vagueness and figurativity

Context and world knowledge should be used only to *disambiguate* the meaning of a word where it actually has multiple senses, not to refine it where it could refer to different things in context. For example, consider the sentences

(1) She felt a sense of shock at the outcome.
(2) She expressed her shock at the outcome.

The word â€˜shockâ€™ is ambiguous: as a technical term it could refer to a mechanical device, or to a medical state, but in the context of (1) and (2) it clearly has a sense corresponding to the FEELING tag.

You might notice that in (2) â€˜shockâ€™ is part of the content of a communication event. However, we do not want to say that â€˜shockâ€™ is ambiguous between an emotional state and something that is communicated; in (2) it is merely a feeling that happens to be communicated, while in (1) it is not communicated. Thus, we do *not* mark it as COMMUNICATION, because this meaning is not inherent to the word itself.

A similar problem arises with metaphor, metonymy, iconicity, and other figurative language. If a building is shaped like a pumpkin, given

(3) She lives in a pumpkin.

you might be tempted to mark â€˜pumpkinâ€™ as an ARTIFACT (because it is a building). But here â€˜pumpkinâ€™ is still referring to the normal sense of pumpkinâ€”i.e. the PLANTâ€”and from context you know that the typical appearance of a pumpkin plant is being used *in a novel (non-standard) way* to describe something that functions as a building. In other words, that buildings can be shaped like pumpkins is not something you would typically associate with the word â€˜pumpkinâ€™ (or, for that matter, any fruit). Similarly, in the sentence

(4) I gave her a toy lion.

â€˜toyâ€™ should be tagged as ARTIFACT and â€˜lionâ€™ as ANIMAL (though it happens to be a nonliving depiction of an animal).

On the other hand, if it is highly conventional to use an expression figuratively, as in (5), we can decide that this figurative meaning has been lexicalized (given its own sense) and tag it as such:

(5) The White House said it would issue its decision on Monday.

According to WordNet, this use of â€˜White Houseâ€™ should be tagged as GROUP (not ARTIFACT) because it is a standard way to refer to the administration.

Highly idiomatic language should be tagged as if it were literal. For example, *road* in the phrase *road to success* should be tagged as ARTIFACT, even if it is being used metaphorically. Similarly, in an expression like

(6) behind the cloak of the Christian religion

(i.e., where someone is concealing their religious beliefs and masquerading as Christian), *cloak* should be tagged as an ARTIFACT despite being used nonliterally.

Supersense classification

Below are some examples of important words in specific domains, followed by a set of general-purpose rules.

Software domain

- pieces of software: COMMUNICATION
  - *version*, *distribution*
  - (software) *system*, *environment*
  - (operating system) *kernel*
- *connection*: RELATION
- *project*: COGNITION
- *support*: COGNITION
- *a configuration*: COGNITION
- *development*: ACT
- *collaboration*: ACT

Sports domain

- *championship*, *tournament*, etc.: EVENT

Science domain

- chemicals, molecules, atoms, and subatomic particles (*nucleus*, *electron*, *particle*, etc.): SUBSTANCE

Other special cases

- *world* should be decided based on context:
  - OBJECT if used like *Earth*/*planet*/*universe*
  - LOCATION if used as a place that something is located
  - GROUP if referring to humanity
  - (possibly other senses as well)
- someone's *life*:
  - TIME if referring to the time period (e.g. *during his life*)
  - STATE if referring to the person's (physical, cognitive, social, ...) existence
  - STATE if referring to the person's physical vitality/condition of being alive
  - (possibly others)
- *reason*: WordNet is kind of confusing here; I think we should say:
  - MOTIVE if referring to a (putative) cause of behavior (e.g. *reason for moving to Europe*)
  - COGNITION if referring to an understanding of what caused some phenomenon (e.g. *reason the sky is blue*)
  - COGNITION if referring to the abstract capacity for thought, or the philosophical notion of rationality
  - STATE if used to contrast reasonableness vs. unreasonableness (e.g. *within reason*)
  - [WordNet also includes COMMUNICATION senses for stated reasons, but I think this is splitting hairs. It makes more sense to contrast MOTIVE/COGNITION vs. COMMUNICATION for *explanation*, where communication seems more central to the lexical meaning. FrameNet seems to agree with this: the Statement frame lists *explanation* but not *reason*.]

Decision list

This list attempts to make more explicit the semantic distinctions between the supersense classes for nouns. Follow the directions in order until an appropriate label is found.

1. If it is a **natural feature** (such as a mountain, valley, river, ocean, cave, continent, planet, the universe, the sky, etc.), label as OBJECT
2. If it is a **man-made structure** (such as a building, room, road, bridge, mine, stage, tent, etc.), label as ARTIFACT
   - includes venues for particular types of activities: *restaurant*, *concert hall*
   - *tomb* and *crypt* (structures) are ARTIFACTS, *cemetery* is a LOCATION
3. For **geopolitical entities** like cities and countries:
   - If it is a **proper name** that can be used to refer to a location, label as LOCATION
   - Otherwise, choose LOCATION or GROUP depending on which is the more salient meaning in context
4. If it describes a **shape** (in the abstract or of an object), label as SHAPE: *hexahedron*, *dip*, *convex shape*, *sine curve* *groove*, *lower bound*, *perimeter*
5. If it otherwise refers to an **space, area, or region** (not specifically requiring a man-made structure or describing a specific natural feature), label as LOCATION: *region*, *outside*, *interior*, *cemetery*, *airspace*
6. If it is a name of a **social group** (national/ethnic/religious/political) that can be made singular and used to refer to an individual, label as PERSON (*Arab*, *Muslim*, *American*, *communist*)
7. If it is a **social movement** (such as a religion, philosophy, or ideology, like *Islam* or *communism*), label as COGNITION if the belief system as a "set of ideas" sense is more salient in context (esp. for academic disciplines like *political science*), or as GROUP if the "set of adherents" is more salient
8. If it refers to an **organization or institution** (including companies, associations, teams, political parties, governmental divisions, etc.), label as GROUP: *U.S. State Department*, *University of California*, *New York Mets*
9. If it is a **common noun** referring to a **type or event of grouping** (e.g., *group*, *nation*, *people*, *meeting*, *flock*, *army*, *a collection*, *series*), label as GROUP

10. If it refers to something being used as **food or drink**, label as FOOD
11. If it refers to a **disease/disorder or physical symptom thereof**, label as STATE: *measles*, *rash*, *fever*, *tumor*, *cardiac arrest*, *plague* (= epidemic disease)
12. If it refers to **the human body or a natural part of the healthy body**, label as BODY: *ligament*, *fingernail*, *nervous system*, *insulin*, *gene*, *hairstyle*
13. If it refers to a **plant or fungus**, label as PLANT: *acorn squash*, *Honduras mahogany*, *genus Lepidobotrys*, *Canada violet*
14. If it refers to a **human or personified being**, label as PERSON: *Persian deity*, *mother*, *kibbutznik*, *firstborn*, *worshiper*, *Roosevelt*, *consumer*, *guardsman*, *glasscutter*, *appellant*
15. If it refers to **non-plant life**, label as ANIMAL: *lizard*, *bacteria*, *virus*, *tentacle*, *egg*
16. If it refers to a category of entity that pertains generically to **all life** (including both plants and animals), label as OTHER: *organism*, *cell*
17. If it refers to a prepared **drug** or health aid, label as ARTIFACT: *painkiller*, *antidepressant*, *ibuprofen*, *vaccine*, *cocaine*
18. If it refers to a **material or substance**, label as SUBSTANCE: *aluminum*, *steel* (= metal alloy), *sand*, *injection* (= solution that is injected), *cardboard*, *DNA*, *atom*, *hydrochloric acid*
19. If it is a term for an **entity that is involved in ownership or payment**, label as POSSESSION: *money*, *coin*, *a payment*, *a loan*, *a purchase* (= thing purchased), *debt* (= amount owed), one's *wealth/property* (= things one owns)
    - Does NOT include *acts* like *transfer*, *acquisition*, *sale*, *purchase*, etc.
20. If it refers to a **physical thing that is necessarily man-made**, label as ARTIFACT: *weapon*, *hat*, *cloth*, *cosmetics*, *perfume* (= scented cosmetic)
21. If it refers to a **nonliving object occurring in nature**, label as OBJECT: *barrier reef*, *nest*, *stepping stone*, *ember*
22. If it refers to a **temporal point, period, amount, or measurement**, label as TIME: *instant/moment*, *10 seconds*, *2011* (year), *2nd millenium BC*, *day*, *season*, *velocity*, *frequency*, *runtime*, *latency/delay*
    - Includes names of holidays: *Christmas*
    - *age* = 'period in history' is a TIME, but *age* = 'number of years something has existed' is an ATTRIBUTE
23. If it is a (non-temporal) **measurement or unit/type of measurement involving a relationship between two or more quantities**, including ordinal numbers not used as fractions, label as RELATION: *ratio*, *quotient*, *exponential function*, *transitivity*, *fortieth/40th*
24. If it is a (non-temporal) **measurement or unit/type of measurement**, including ordinal numbers and fractional amounts, label as QUANTITY: *7 centimeters*, *half*, *1.8 million*, *volume* (= spatial extent), *volt*, *real number*, *square root*, *decimal*, *digit*, *180 degrees*, *12 percent/12%*
25. If it refers to an **emotion**, label as FEELING: *indignation*, *joy*, *eagerness*
26. If it refers to an **abstract external force that causes someone to intend to do something**, label as MOTIVE: *reason*, *incentive*, *urge*, *conscience*
    - NOT *purpose*, *goal*, *intention*, *desire*, or *plan*
27. If it refers to a person's **belief/idea or mental state/process**, label as COGNITION: *knowledge*, *a dream*, *consciousness*, *puzzlement*, *skepticism*, *reasoning*, *logic*, *intuition*, *inspiration*, *muscle memory*, *theory*
28. If it refers to a **technique or ability**, including forms of perception, label as COGNITION: *a skill*, *aptitude/talent*, *a method*, *perception*, *visual perception/sight*, *sense of touch*, *awareness*
29. If it refers to an act of **information encoding/transmission** or the abstract information/work that is encoded/transmittedâ€”including the use of language, writing, music, performance, print/visual/electronic media, or other form of signalingâ€—label as COMMUNICATION: *a lie*, *a broadcast*, *a contract*, *a concert*, *a code*, *an alphabet*, *an equation*, *a denial*, *discussion*, *sarcasm*, *concerto*, *television program*, *software*, *input* (= signal)
    - Products or tools facilitating communication, such as books, paintings, photographs, or televisions, are themselves ARTIFACTS when used in the physical sense.
30. If it refers to a **learned profession** (in the context of practicing that profession), label as ACT: *engineering*, *law*, *medicine*, etc.
31. If it refers to a **field or branch of study** (in the sciences, humanities, etc.), label as COGNITION: *science*, *art history*, *nuclear engineering*, *medicine* (= medical science)
32. If it refers in the abstract to a **philosophical viewpoint**, label as COGNITION: *socialism*, *Marxism*, *democracy*
33. If it refers to a **physical force**, label as PHENOMENON: *gravity*, *electricity*, *pressure*, *suction*, *radiation*
34. If it refers to a **state of affairs**, i.e. a condition existing at a given point in time (with respect to some person/thing/situation), label as STATE: *poverty*, *infamy*, *opulence*, *hunger*, *opportunity*, *disease*, *darkness* (= lack of light)
    - heuristic: in English, can you say someone/something is "in (a state of) X" or "is full of X"?

- let's exclude anything that can be an emotion [though WordNet also lists a STATE sense of *happiness* and *depression*]
- easily confused with ATTRIBUTE and FEELING

35. If it refers to an **aspect/characteristic of something that can be judged** (especially nouns derived from adjectives), label as ATTRIBUTE: *faithfulness*, *clarity*, *temperature* (= degree of hotness), *valence*, *virtue*, *simplicity*, *darkness* (= dark coloring)
    - easily confused with STATE, FEELING, COGNITION
36. If it refers to the **relationship between two entities**, label as RELATION: *connection*, *marital relationship*, (non-person) *member*, (non-statistical) *correlation*, *antithesis*, *inverse*, *doctor-patient relationship*, *politics* (= social means of acquiring power), *causality*
37. If it refers to **"something that people do or cause to happen"**, label as ACT: *football game*, *acquisition* (= act of acquiring), *hiring*, *scoring*
    - Includes wars.
38. If it refers to **"something that happens at a given place and time"** label as EVENT: *tide*, *eclipse*, *accident*
    - Includes recurring events like sports tournaments.
39. If it refers to **"a sustained phenomenon or one marked by gradual changes through a series of states"** (esp. where the changes occur in a single direction), label as PROCESS: *evaporation*, *aging*, *extinction*, (economic) *inflation*, *accretion*/*growth*
40. If it refers to **something that happens/occurs**, label as PHENOMENON: *hurricane*, *tornado*, *cold front*, *effect*
41. If it is a synonym of ***kind/variety/type* (of something)**, label as COGNITION
42. If it is part of a **stock phrase used in discourse**, label as COMMUNICATION: for *example*, on the one *hand*, in the *case* of
43. If it is some other **abstract concept that can be known**, it should probably be labeled as COGNITION.

If you cannot decide based on these guidelines, use the "UNSURE" tag.