Basic Natural Language Terminology

Sangkeun Jung

Content

- ✓ NLP Terminology

TEXT BASED SEGMENTATION

Structure | Paragraph

In the Extended Lexicon, we introduce a third kind of linguistic object, called word instances (or just instances), consisting of word forms as they occur in strings (sequences of words, typically sentences). For example, a string such as *the cats sat on the mat* contains two distinct instances of the word *the*. *the cats slept* contains further (distinct) instances of *the* and *cats*. However the instances in a repetition of *the cats sat on the mat* are the same as those in the original (because instances are defined relative to strings, that is, string types not string tokens).

So in an extended lexicon, the lexical entries are word instances, and the lexicon itself is a structured description of a set of word instances. In order to explore this notion in more detail, it is helpful to introduce a more specific notion of a 'structured description'. We shall use an inheritance-based lexicon, in which there are internal abstract

Structure | Paragraph | Sentence

In the Extended Lexicon, we introduce a third kind of linguistic object, called word instances (or just instances), consisting of word forms as they occur in strings (sequences of words, typically sentences). For example, a string such as *the cats sat on the mat* contains two distinct instances of the word *the*. *the cats slept* contains further (distinct) instances of *the* and *cats*. However the instances in a repetition of *the cats sat on the mat* are the same as those in the original (because instances are defined relative to strings, that is, string types not string tokens).

So in an extended lexicon, the lexical entries are word instances, and the lexicon itself is a structured description of a set of word instances. In order to explore this notion in more detail, it is helpful to introduce a more specific notion of a 'structured description'. We shall use an inheritance-based lexicon, in which there are internal abstract

In the Extended Lexicon, we introduce a third kind of linguistic object, …

In linguistics, a **word** is the smallest element that can be uttered in isolation with objective or practical meaning.
This contrasts deeply with a morpheme, which is the smallest unit of meaning but will not necessarily stand on its own. A word may consist of a single morpheme (for example: *oh!, rock, red, quick, run, expect*), or several (*rocks, redness, quickly, running, unexpected*), whereas a morpheme may not be able to stand on its own as a word (in the words just mentioned, these are *-s, -ness, -ly, -ing, un-, -ed*).

(from Wikipedia)

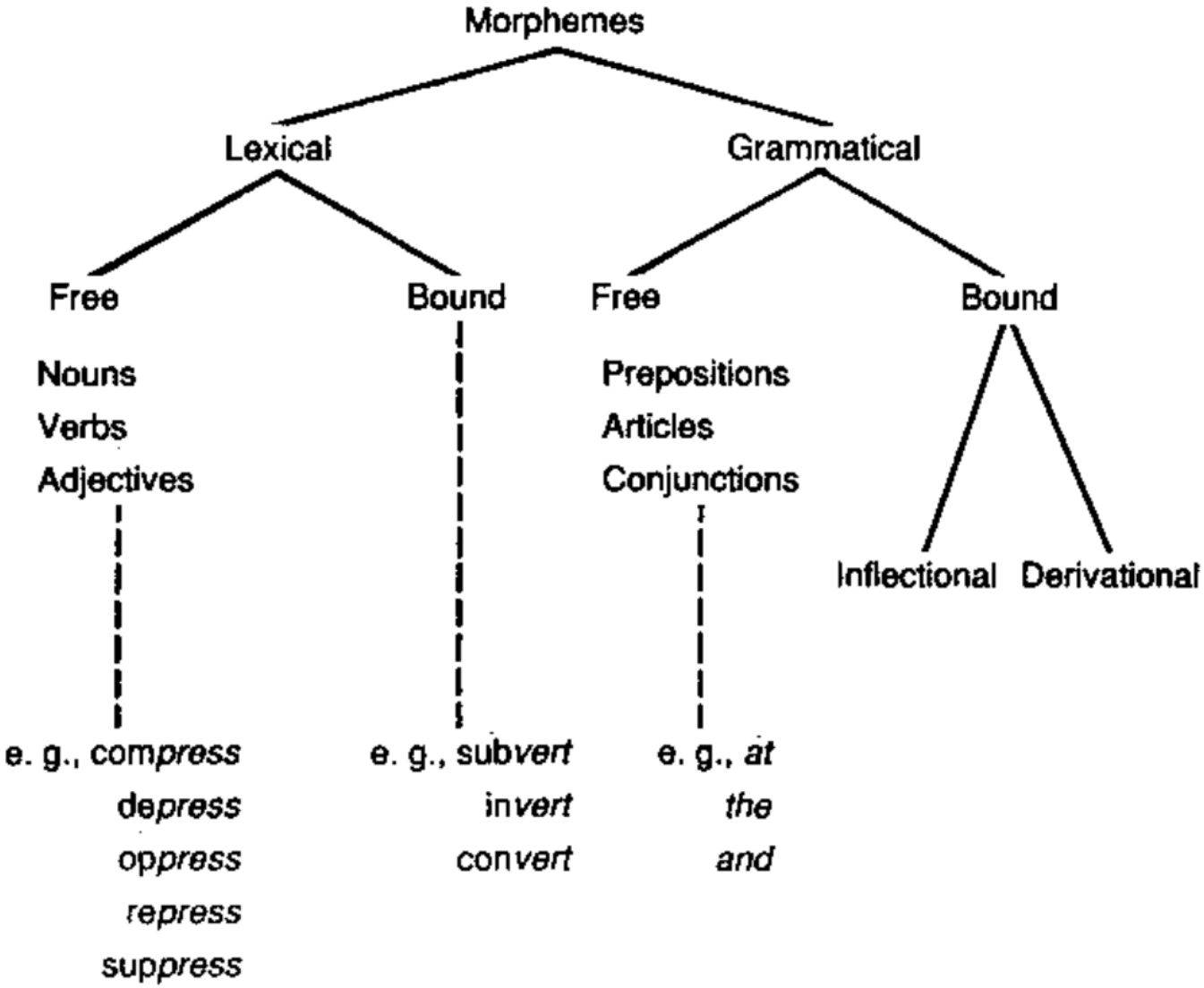
Extended ➔ extend + ~ed

Root morpheme

Affix morpheme

A morpheme is the **smallest grammatical unit** in a language. A morpheme is not identical to a word, and the principal difference between the two is that a morpheme may or may not stand alone, whereas a word, by definition, is freestanding. (from Wikipedia)

Structure | Paragraph | Sentence | Words | Morphemes



http://www.mathcs.duq.edu/~packer/Courses/Psy598/Ling-Morphology.pdf

Structure | Paragraph | Sentence | Words | Morphemes (Lexical, Grammatical)

- ✓ Lexical morphemes are those that having meaning by themselves (more accurately, they have sense).
- ✓ Grammatical morphemes specify a relationship between other morphemes.
- ✓ But the distinction is not all that well defined.

- ✓ Nouns, verbs, adjectives ({boy}, {buy}, {big}) are typical lexical morphemes.
- ✓ Prepositions, articles, conjunctions ({of}, {the}, {but}) are grammatical morphemes.

http://www.mathcs.duq.edu/~packer/Courses/Psy598/Ling-Morphology.pdf

Structure | Paragraph | Sentence | Words | Morphemes (Free, Bound)

- ✓ ***Free morphemes*** are those that can stand alone as words. They may be lexical morphemes ({serve}, {press}), or grammatical morphemes ({at}, {and}).

- ✓ ***Bound morphemes*** can occur only in combination—they are parts of a word. They may be lexical morphemes (such as {*clude*} as in include, *exclude*, *preclude*) or they may be grammatical (such as {PLU} = plural as in boys, girls, and cats).

http://www.mathcs.duq.edu/~packer/Courses/Psy598/Ling-Morphology.pdf

Structure | Paragraph | Sentence | Words | Grapheme(자소)

In linguistics, a **grapheme** is the smallest unit of a writing system of any given language.[1] An individual grapheme may or may not carry meaning by itself, and may or may not correspond to a single phoneme of the spoken language. Graphemes include alphabetic letters, typographic ligatures, Chinese characters, numerical digits, punctuation marks, and other individual symbols. A grapheme can also be construed as a graphical sign that independently represents a portion of linguistic material.[2]

(from Wikipedia)

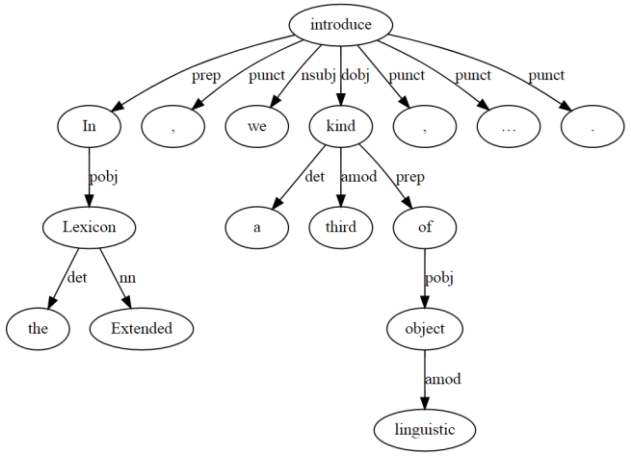
http://www.readingdoctor.com.au/phonemes-graphemes-letters-word-burger

Sky ➔ s, k, y

天空 ➔ 天, 空

LETTERS are the visual building blocks of written words. When we say the letters in a word, as in C (see) A (ay) T (tee), we are describing the way the word looks, not the way it sounds.

Lexical, Orthography, Syntax



… *Syntactic Features*

In the Extended Lexicon, we introduce a third kind of linguistic object, …

… *Orthography Features*

(철자법, 맞춤법)

An orthography is a set of conventions for writing a language. It includes norms of spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation.

In the Extended Lexicon, we introduce a third kind of linguistic object, …

… *Lexical Features*

SPEECH BASED SEGMENTATION

Utterance (발화)

In spoken language analysis, an **utterance** is the smallest unit of speech. It is a continuous piece of speech beginning and ending with **a clear pause**.

In the case of oral languages, it is generally but not always bounded by silence. Utterances do not exist in written language, only their representations do. They can be represented and delineated in written language in many ways.

(from Wikipedia)

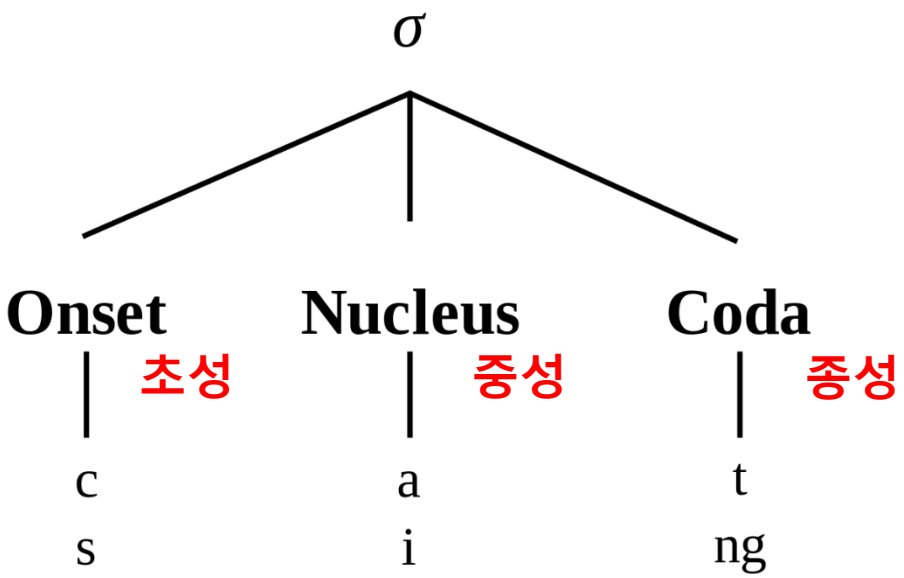
> The main difference between **sentence** and **utterance** is that the sentence coveys a complete meaning, either spoken or written, whereas utterance usually does not necessarily convey a complete meaning.
>
> Communication is the only way two human beings can interact and share their thoughts and sentiments with each other. There are two major means of communication as verbal and non-verbal communication. **Sentences**, therefore, belong to both verbal and non-verbal types of communication since they can either be spoken or written. Yet an **utterance** is typically a sound or incomplete spoken group of words that belong to the verbal type of communication.

http://pediaa.com/difference-between-sentence-and-utterance/

Syllable (음절), Consonant(자음), Vowel(모음)

A **syllable** is a unit of organization for a sequence of speech sounds. It is typically made up of a syllable nucleus (most often a vowel) with optional initial and final margins (typically, consonants). Syllables are often considered the phonological "building blocks" of words.[1]

(from Wikipedia)



Segmental model for *cat* and *sing*

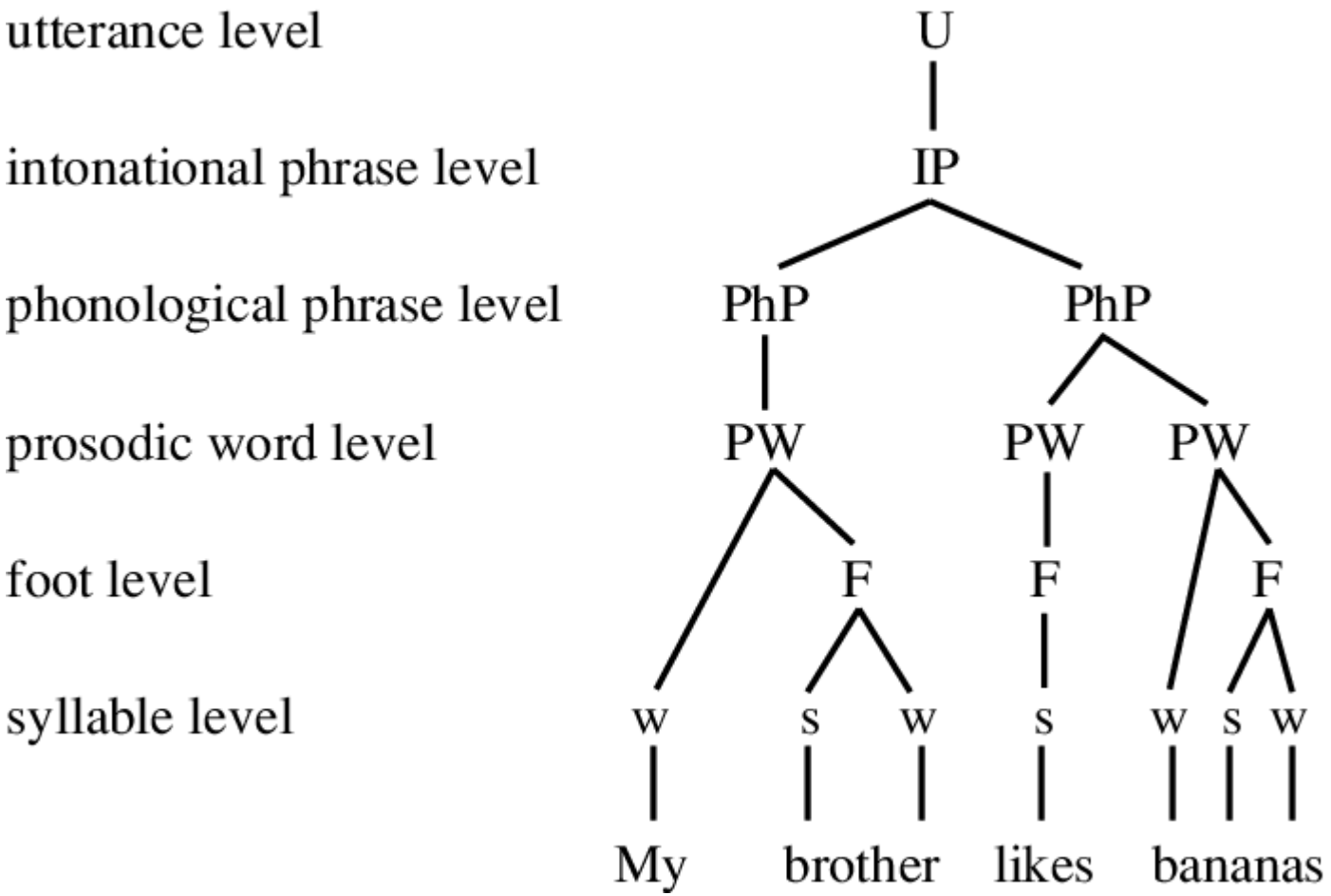
Phoneme (음소)

In linguistics, phonemes (usually established by the use of minimal pairs, such as *kill* vs *kiss* or *pat* vs *bat*) are written between slashes, e.g. /p/. To show pronunciation more precisely linguists use square brackets, for example [p^h] (indicating an aspirated *p*). (from Wikipedia)

Phonemes are speech sounds made by the mouth, like the /p/ sound in /spoon/.

http://www.readingdoctor.com.au/phonemes-graphemes-letters-word-burger

Prosodic Hierarchy (운율체계)



[An example of the prosodic hierarchy.]

https://www.researchgate.net/publication/273037379_An_Overview_of_Prosody_and_Its_Role_in_Normal_and_Disordered_Child_Language

(Application)

- Text based segmentation to Speech based Segmentation is the key of developing “Text to Speech”.