Conversational Lexical Standards

Kurt Fuqua, Cambridge Mobile
kurt@cambridgemobile.com

Abstract Intelligent morphology processing is essential to creating conversational applications and engines. This intelligence allows developers to create more powerful apps with less need to understand linguistics. A single standardized lexicon can be shared among engines and apps and the process becomes more efficient. This paper discusses standardization of parts of speech, grammatical features, phoneme set, morphologic processing, and a formalism for defining lexical grammars.

Are the words 'apple' and 'apples' related? Despite the obvious answer, many systems treat these word forms as if they are unrelated semantically and morphologically. This paper addresses the need for lexical standards in order to support conversational applications.

Conversation requires tracking linguistic information across multiple applications. A virtual agent is constantly tracking pronouns, and anaphoric references, expanding ellipsis and sending the expanded, resolved statements to the appropriate application. This requires a much higher level of abstraction than required in stand-alone, non-conversational applications. The applications must share a common syntactic and lexical grammar, as well as a common lexicon. The information shared between the applications must contain high-level semantic information.

The Need for Morphology

If an application is to be conversational, it must support morphology. These are the changes that words undergo during regular inflection. The conversational system should be able to process these regular changes in an intelligent, symbolic fashion. An application developer may need to synthesize a statement such as:

"You have three messages."

The pluralization of 'message' should be done symbolically with automatic agreement between the number and noun. The lexicon should only need to store the root form of each word; a morphologic engine can synthesize the proper plural form for the given language. The identical lexicon should be used for both synthesis and analysis. If there are two different lexicons (one for synthesis and another for analysis) then synchronization problems arise. Likewise the morphologic engine should be capable of processing morphology in both the synthesis and analysis directions.

From the root form, the engine can derive every regularly-derived form. This assures consistent coverage. In current systems, coverage is inconsistent. It may be possible to use a verb in the past tense but not the future tense, or a noun in the singular but not the plural.

There are 3 required components:

- The lexicon information,
- A formalism for defining the grammar,
- The morphologic engine for processing.

What is in a Lexicon?

There are two main distinctions between a word list and a lexicon. A lexicon contains grammatical information – not just pronunciations. Second, lexicons are lemmatized – that is they contain the root form of the words. A lexicon is the repository for the grammatical information on each lemma (root form).

Each lexicon entry should contain:

- The graphemic representation: 'message'
- The phonemic representation: /mɛsəʤ/
- The part(s) of speech: noun
- The grammatical feature-values

- Language-specific token
- Semantic token

Each of these elements must be standardized and be consistent between the lexical grammar, morphologic engine, and application. Moreover, this consistency should extend across languages to the extent possible.

Parts of Speech

The most basic grammatical information is the part of speech (POS). Very little processing can be done without knowing a word's part of speech. Currently, there is no W3C standard for POS. This is the most basic requirement for creating a lexicon standard.

While word categories are not without controversy, it is possible to construct a set of basic POS tags which are indeed universal. This proposal follows the SLAPI POS standard. This standard has been used for dozens of languages.

Open Class POS (5)
verb, adverb, adjective, common noun, proper noun

Closed Class POS (10)
pronoun, number, conjunction, preposition, determiner, quantifier, interjection, portmaneux, clitic, punctuation

Binary representation of the POS is extremely important to efficient implementations. With 15 POS, the POS of each polysemic entry can be represented in bitwise form with just 16 bits. The importance of the bitwise representation cannot be overemphasized! Therefore it is necessary to limit the POS to 15 categories.

Features

Each part of speech is subcategorized with grammatical features. Examples include:

Plurality	plur
Count	count
Case	case
Animate	animate
Natural Gender	ngender
Grammatical Gender	ggender
Mood	mood
Valency	valency
Pronoun Type	prontype
Conjunction type	conjtype

Each feature has possible values. Some are normally Boolean with a value of + or -. Others have specifically enumerated values. These features can be used to tag entries in the lexicon or to derive word forms. The word qualifier is placed in brackets after the word.

'message' <plur=+>

This has the meaning 'the plural form of the word message'. The developer does not need to concern himself with how the morphology of English works, or whether that word is a regular forming plural – the morphology engine will inflect it properly according to the defined lexical grammar.

While the aggregate potential values for a given feature are fixed, the specific possible values may vary from language to language.

English
n plur = (+, -)

Arabic
n plur = (+,-, dual)

In a report titled "*Towards a Standard for the Creation of Lexica*", a group of researchers set out a harmonized feature set for twelve (12) European languages. The current SLAPI feature set encompasses all of these features plus several required for other languages. About 175 grammatical features with their possible values are defined. This is a good start but there is room for improvement.

Pronunciations

Pronunciations are only meaningful if they can be used across applications and tools in a

consistent way. There is currently gross inconsistency between vendors and even within a single vendor. The first requirement is to agree on what is represented. Does the pronunciation symbol represent a phoneme or an allophone? The W3C standards are deliberately ambiguous. Some vendors represent phonemes, others allophones. One widely used mobile phone product from Google actually contains an admixture of *both* phonemes and allophones. Such products are unworkable even for a trained computational linguist.

For standardized lexicons the symbols must be phonemes – not allophones! Phonologic rules transform phonemes into allophones. There are regular phonologic rules which apply within a word to morphologic derivations. Other Sandhi rules apply between adjacent words. Without phonemic symbols, phonologic rules cannot be applied.

If the lexicons always contain phonemic entries, a common engine can apply phonologic rules in a systematic, efficient and accurate manner.

Phoneme Sets

Every language has a well-defined set of phonemes from which all words of the language are constructed. This phoneme set is the spoken equivalent of the written alphabet. Determining a language's phoneme set is fairly objective for a linguist. The standard should include a published set of defined phonemes for each language, and the symbols to represent those phonemes.

Even when scholars agree on the set of phonemes, and use IPA to represent those phonemes, they may use different symbols. Rarely do 2 dictionaries use the same symbols.

The Cambridge Phoneme Set (CPS) defines the phonemes for about 52 languages. In addition to defining the phonemes used in each language, it defines a consistent set of symbols for representing those phonemes. The principals are described in the document. This is a good starting point for a lexical standard.

In order to assure adequate minimal processing, every engine should be required to support a defined minimal set of phonemes. SLAPI and CPS define a required set of 42 consonantal phonemes. This includes every phoneme required for English, German, Danish, Dutch, Norwegian, Farsi, Japanese, French, Italian, Spanish, Portuguese, Romanian, Catalan, Finnish, Lithuanian, Slovenian, and Bulgarian.

Core Lexicon Certification

A common lexicon must be shared among the conversational applications. Irrespective of the topic, there is a certain vocabulary content that ought to be included for any conversational application. This is referred to as the 'core'. The core should contain all closed parts of speech (i.e. pronouns, prepositions, conjunctions, determiners, etc), plus the most common words from the open parts of speech (verbs, nouns, adjectives, adverbs, proper nouns) and the most common irregular forms. To this core, additional vocabulary can be added.

Generally, it is the lexicon core which involves the greatest development effort. The core often involves a great deal of polysemy and exceptions, and must be manually created whereas the remainder of the lexicon can be created through automated or semi-automated processes.

The standard should contain metrics for measuring the completeness of the core. Under SLAPI, there are two levels defined: Core1 and Core2. To qualify as Core1 complete, the lexicon must meet numeric completeness, feature completeness and semantic completeness. Here is a partial list of the criteria for a Core1 complete lexicon. It must contain at least: all non-archaic closed-part-of-speech entries, 500 lexical lemma entries, 100 distinct verbs, 100 common nouns, 50 adj, 50 adv, and 10 part-of-speech-polysemic entries,

all days of the week, months and seasons, all cardinals, ordinals, six basic colors, 15 temporal adverbs, 15 spatial adverbs, 9 day period nouns.

There are also criteria to measure lexical grammar completeness, and semantic commonality between pairs of language lexicons. This assures that semantic coverage can be objectively compared.

Semantic Tokens

Semantic representation has always been controversial but the potential benefit is high. For cooperative, conversational applications, semantics are required. InterLing is a well-defined interlingua based upon predicate logic. It has a reasonably large commercial vocabulary. Lexicon entries can be tagged with the appropriate InterLing term. This semantic information can then be a mechanism for sharing semantics across applications. Instead of passing language-specific strings, simple language-independent tokens are passed. This opens the door to creating language-independent applications.

Validation

Lexicons are notoriously prone to contain errors. However, there are some simple additions which can allow partial validation of lexicons automatically. For example, an orthographic alphabet for the language should be defined. Any entry containing an orthographic symbol outside of that set can be caught. Similarly, the defined phoneme set will allow detection of impossible pronunciations. If transliteration rules (G-to-P) are included, then any entry not conforming to the expected pronunciation, and not tagged as exceptional, will be caught. Phonotactic rules allow for further checks of syllabic structure.

Polysemy

Since polysemy is pervasive in natural-language conversation, a conversational lexicon must adequately represent polysemic entries. Polysemy can be divided into 2 levels. Part-of-speech polysemy refers to a word with multiple possible parts-of-speech. The word 'plan' is both a verb and a common noun. Feature polysemy refers to a word that has more than meaning within a single part of speech. For example the word 'her' is both a personal pronoun and a possessive pronoun.

Formalism for Lexical Grammar

The term grammar for speech systems generally connotes syntactic grammar. There is also a need for a formalism to create lexical grammars. Such a grammar should define:

- Graphemic alphabet
- Phoneme set
- Phonotactics
- Phonology rules
- Prosody
- Transliteration rules
- Applicable grammatical features for each POS
- Morphology (all allomorphs)

SLAPI uses a formalism called Lingua which was originally developed at IIT in the 1980's. Lingua represents phonology based upon the widely accepted Optimality Theory.

Morphology Engine

Ideally, the morphology engine would be universal, that is, the same engine could be used for multiple languages. If the engine is language-independent, it can be completely driven by the lexical grammar and the data from the lexicon. This would allow applications to potentially change the natural language without an architectural change.

A universal engine would need to process many of the common morphologic processes found in dominant languages. An abbreviated list includes: allomorphs, suppletive allomorphs, concatenation rules for allomorphs, suffixes, prefixes, circumfixes, infixes, paraphrastics, and multilevel allomorphic precedence.

The engine must be bidirectional, handling both directions of the conversation (analysis and composition). It must also be equally capable of processing phonemic and graphemic forms.

Script processing

Much of NL processing has been biased toward the Roman (Latin) script. The emergence of the ISO 15924 standard has given increased visibility to the other ~150 NL scripts. Many of these scripts function very differently. Abugidas, in particular, are quite complex. Most scripts are not bicameral. Some contain allographs or systemic ligatures. Many contain special digits. Some use different word delimiters or no delimiters (scriptio continua). Graphemic processing should be delegated to the morphologic engine as well. To this end, the scripts themselves should also be systematically categorized. SLAPI defines 11 script features.

Conclusion

It is essential to standardize part-of-speech tokens, tokens for grammatical features, and phoneme sets so that lexicons and lexical information can be shared across applications. The higher level of abstraction enables conversationally enabled applications to cooperate. A lexical grammar would define the morphology and phonology. This would be implemented in a distinct morphology engine which would process spoken and textual natural language. In order to enable conversation, the engine would be driven by the lexical grammar and function both in the analysis and synthesis directions.

Appendix

Parts of Speech
Open Class POS (5)
verb, adverb, adjective, common noun, proper noun

Closed Class POS (10)
pronoun, number, conjunction, preposition (& postpositionals),
determiner, quantifier, interjection, portmaneux (& contractions),
clitic (& particles), punctuation (& symbols)

Features
A few common features taken from English or French:

```
plur   =binary                                        ! plural, singular
count  =binary                                        ! count/mass noun
pernum =(p1,p2,p3,p4,p5,p6)                           ! person & number combined
time   =(pres,past,fut)                               ! cf Aux tense for v
case   =(nom,dat,acc)                                 ! nominative/dative case
gen    =binary                                        ! genitive
negative=binary
admode =(spatial,temporal,degree,process,contingency,modality,manner,purpose)
temptype=(tpoint,tdur,tfreq,trel)                     ! adj & adv temporal type
spactype=(spos,sdir)                                  ! spatial type
gradable=binary                                       ! adj & adv
form   =(absolute,compar,super)                       ! adj & adv superlative
interg =binary                                        ! interrogative
def    =binary                                        ! definate
animate =binary                                       ! animate
valency=(modal,nultrans,intrans,trans,ditrans,copulaA,copulaN,copulaV)
mood   =(inf,indicative,prespt,pastpt,imperative,subjunctive)
reflexive=binary
conjugation=(first,second,third,fourth)
ggender     =(masc,fem)                               ! grammatical gender
```

Lexicon Entries
Example lexicon entries in symbolic form

‘message’ /mɛsəʤ/	noun	<count=+; regplur=+; abstract=+>
‘sing’ /sɪŋ/	verb	<valency=(intrans,trans); regmood=-; mood=(indic,imper,inf,subj)>
‘sung’ /sʌŋ/	verb	<root=’sing’; mood=pastpart>
‘hope’ /ho͡ʊp/	noun|verb	<noun: count=+; regplur=+; abstract=+; verb: valency=intrans>
‘an’ /æn/	det	<count=-; def=-; plural=-; aliteration=+>

Interpretation
The word ‘message’ is an abstract, count noun which forms plurals regularly.

The word ‘sing’ can be used as either an intransitive or transitive verb, is irregular with respect to mood, and this form represents the indicative, imperative, infinitive and subjunctive forms.

The word ‘sung’ is the past participle of the verb root ‘sing’. (Features of the root are inherited.)

The word ‘hope’ is both a noun and verb. It is an abstract, count noun which forms plurals regularly.

The word ‘an’ is a determiner. It is a non-count, indefinite, singular for alliterated noun phrases.

Bibliography

Towards a Standard for the Creation of Lexica, May 2003, Monachini, et al

The Scalable Language IPA, Technical Reference Manual, January 2009

Cambridge Phoneme Set, Reference Manual, 6.0c