Introduction to Computational Linguistics

Michael P. Oakes

University of Wolverhampton.

Michael.Oakes@wlv.ac.uk

Speech and Language Processing

Daniel Jurafsky and James H. Martin,

Prentice Hall, First Edition.

Introduction

- *Dave Bowman: Open the pod bay doors, HAL.*
- *HAL: I'm sorry Dave, I'm afraid I can't do that.*
- From the Screenplay of 2001: A Space Odyssey.
- What would it take to create at least the language-related parts of HAL?
- Understanding humans via **speech recognition** and **natural language understanding** (and, of course **lip-reading**), and of communicating with humans via **natural language generation** and **speech synthesis**. HAL would also need to be able to do **information retrieval** (finding out where needed textual resources reside), **information extraction** (extracting pertinent facts from those textual resources) and **inference** (drawing conclusions based on known facts).

Language Processing Systems

- Language processing systems range from mundane applications such as **word counting** to **spelling correction**, **grammar checking**, and cutting edge applications such as **automated question answering** on the web and **real-time spoken language translation.**
- What distinguishes them from other data processing systems is their use of *knowledge of language.* Even the unix word count program (wc) has knowledge of what constitutes a word.

Levels of Language (1)

- To determine what Dave is saying, HAL must be able to analyse the incoming audio signal. Similarly HAL must be able to generate an audio signal that Dave can understand. These tasks require knowledge of **phonetics** (how words are pronounced in terms of individual speech units called **phones,** listed in the international phonetic alphabet) and **phonology** (the systematic way that sounds are differently realised in different environments, e.g. *cat, cook*).
- HAL is capable of producing contractions like I'm and can't. Producing and recognizing these and other variations of individual words (e.g. recognising that doors is plural) requires knowledge of **morphology.**

Levels of Language (2)

- HAL has knowledge of **syntax**, rules for the combination of words. He knows that the sequence *I'm I do, sorry that afraid Dave I'm can't* will not make sense to Dave, even though it contains exactly the same words as the original. He has knowledge of **lexical semantics** (the meanings of words, e.g. the difference between *door* and *window*, *open* and *shut*. Compare with compile time and run time errors in a computer program.
- Next, despite it's bad behaviour, HAL knows enough to be polite to Dave, embellishing his responses with I'm sorry and I'm afraid. The appropriate use of polite and indirect language comes under **pragmatics**. HAL's correct use of the word *that* in its response to Dave's request provides structure in their conversation, which requires knowledge of discourse conventions.
- See "The scope of linguistics", p14 of "Teach Yourself Linguistics" by Jean Aitchison.

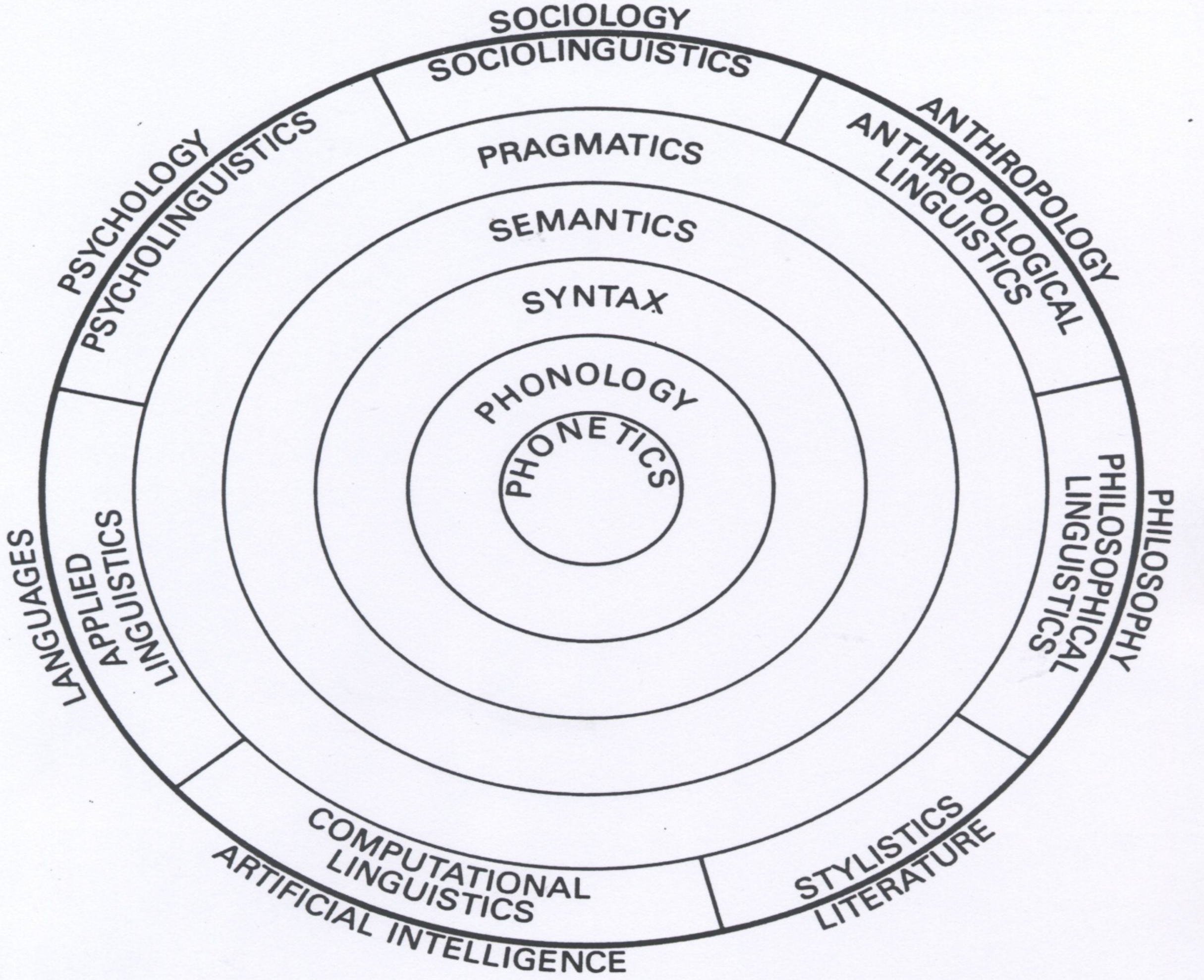

SOCIOLOGY
SOCIOLINGUISTICS
ANTHROPOLOGY
ANTHROPOLOGICAL LINGUISTICS
PHILOSOPHY
PHILOSOPHICAL LINGUISTICS
STYLISTICS
LITERATURE
COMPUTATIONAL LINGUISTICS
ARTIFICIAL INTELLIGENCE
LANGUAGES
APPLIED LINGUISTICS
PSYCHOLOGY
PSYCHOLINGUISTICS
PRAGMATICS
SEMANTICS
SYNTAX
PHONOLOGY
PHONETICS

Fig. 1

To summarize, the knowledge of language needed to engage in complex behaviour can be separated into six distinct categories:

- Phonetics and Phonology - the study of linguistic sounds
- Morphology - the study of the meaningful components of words
- Syntax - the study of the structural relationships between words
- Semantics - the study of meaning
- Pragmatics - the study of how language is used to accomplish goals
- Discourse - the study of linguistic units larger than a single utterance
- Most or all tasks in speech and language processing can be viewed as **resolving ambiguity** at one of these six levels. How many different meanings can you think of for the sentence *I made her duck?*

Resolving Ambiguity

- *Duck* can be a verb or noun, while *her* can be a dative pronoun or a possessive pronoun. *Make* can mean *create* or *cook* or *compel*. It can also be transitive, taking a single direct object (I cooked waterfowl belonging to her) or ditransitive, taking two objects, meaning that the first object (*her*) was made into the second object (*duck*). In a spoken sentence, there would be another kind of ambiguity. What is it?
- How do we **resolve** or **disambiguate** these ambiguities?
- deciding whether *duck* is a verb or a noun can be solved by **part-of-speech tagging**
- deciding whether *make* means *create, cook* or *compel* can be solved by **word sense disambiguation.**
- Deciding whether *make* is transitive or ditransitive is an example of **syntactic disambiguation** and can be addressed by **probabilistic parsing**.

Regular Expressions

- The **regular expression** is the standard notation for characterising text sequences. Useful for web searching, computation of frequencies in corpora, etc. Regular expressions can be implemented via the **finite state automaton**. They are key components of the **Perl** programming language.
- Regular expression search requires a **pattern** that we want to search for, and a **corpus** of texts to search through. A regular expression search function will search through the corpus returning all texts that contain the pattern.
- e.g. /[Ww]oodchuck/ matches *Woodchuck* or *woodchuck*,
- /[A-Z]/ matches any upper case letter,
- /colo?r/ matches *colour* or *color*,
- /\bthe\b/ matches *the* but not *mother*, *thespian* or *absinthe*
- /gupp(y|ies)/ matches *guppy* or *guppies*
- /baaa*!/ uses the Kleene star to define sheep language - *baa!, baaa!, baaaa!, baaaaa! ...*

Substitutions

- An important use of regular expressions is in substitutions.
- e.g. s/colour/color/ replaces *colour* with *color* every time it is found in the text.
- Substitutions may use memory
- e.g. /the (.*)er they were, the \1er they will be/
- will match *the bigger they were, the bigger they will be* but not *the bigger they were, the faster they will be.*
- Substitutions using memory are very useful in implementing a simple natural-language understanding program like ELIZA (Weizenbaum, 1966), which could carry on conversations like the following:

ELIZA

- User: Men are all alike.
- ELIZA: IN WHAT WAY
- User: They’re always bugging us about something or other.
- ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE
- User: Well, my boyfriend made me come here.
- ELIZA: YOUR BOYFRIEND MADE YOU COME HERE
- User: He says I’m depressed much of the time.
- ELIZA: I AM SORRY TO HEAR THAT YOU ARE DEPRESSED.

- ELIZA works by searching the user’s sentence for regular expressions and substituting them, e.g. s/my/YOUR/ and s/I’m/YOU ARE/, and then:
- s/.* YOU ARE (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/
- s/.* always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/

Finite State Automata

- **Finite state automata** are the theoretical foundation for much of the work in NLP. Any regular expression can be implemented as a finite state automaton (FSA) except those which use the memory feature; any FSA can be described with a regular expression. Both FSA and regular expressions can be used to describe a **regular language.**
- See pages 34 – 35 (next slide), which define “sheep language”, which is any word from the following set:
- baa! baaa! baaaa! baaaaa! baaaaaa!...

Finite state automaton

- The FSA has 5 states, q0 to q4
- q0 is the start state, q4 is the end state
- States are nodes, transitions are arcs
- If we are in the final state when we have just read the last input symbol, the input is accepted.
- If we get stuck in some non-final state, the input is rejected.

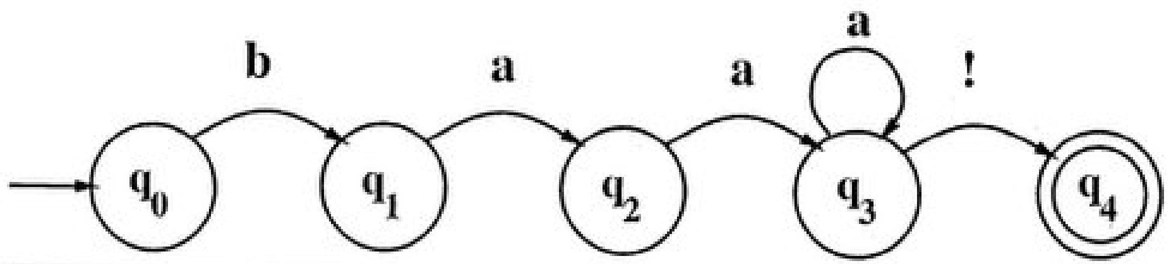


Figure 2.10 A finite-state automaton for talking sheep.

	Input		
State	b	a	!
0	1	∅	∅
1	∅	2	∅
2	∅	3	∅
3	∅	3	4
4:	∅	∅	∅

Figure 2.12 The state-transition table for the FSA of Figure 2.10.

Exercises

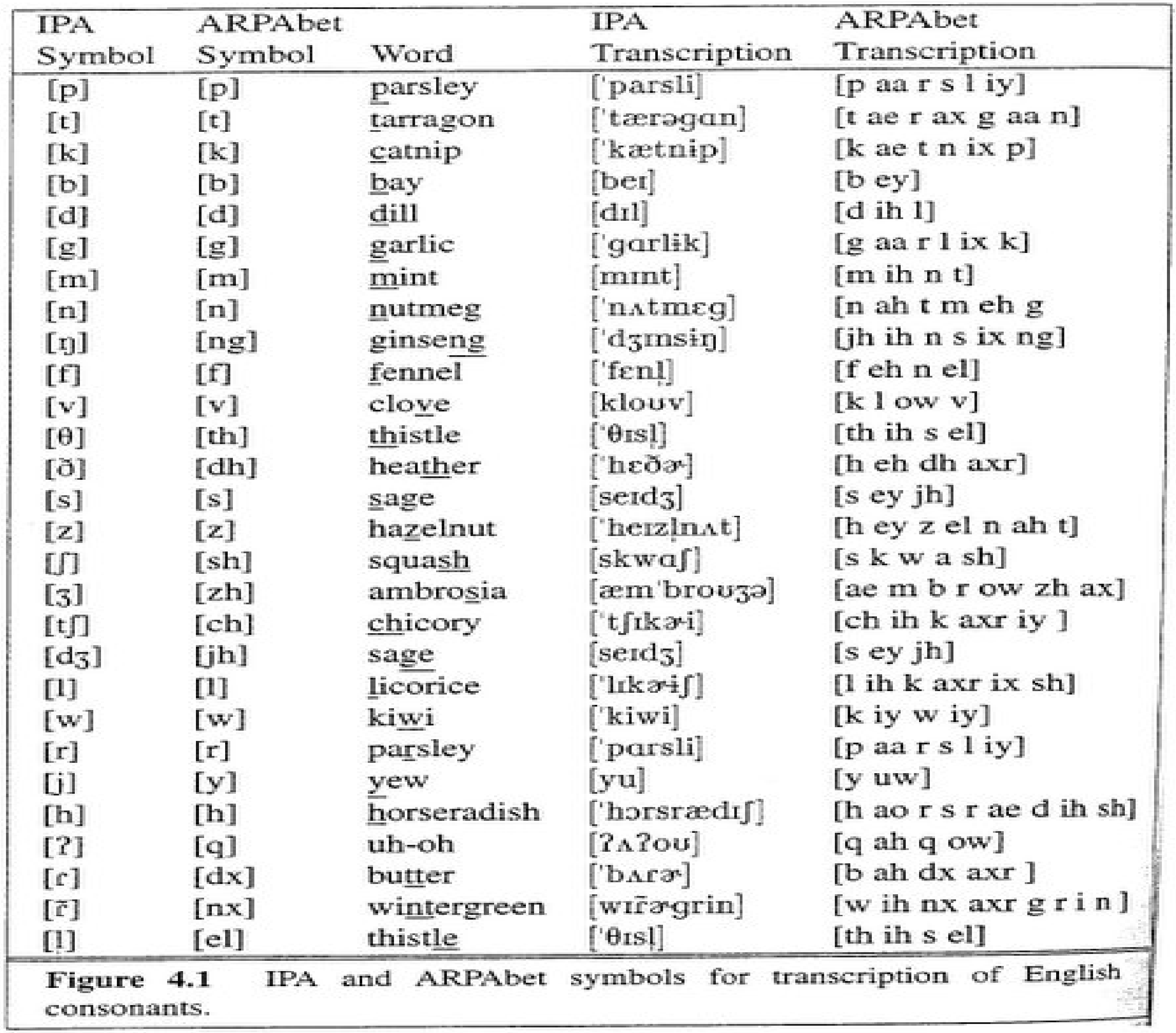
- Exercise: Design an FSA for the words for English numbers 1 to 99.
- Exercise: write Perl-like regular substitution expressions to produce a **finite state transducer** which converts all regular English nouns into their plural forms (**inflectional morphology**)

In order to build a morphological parser, we need at least the following:

- 1. **Lexicon:** the list of stems and affixes, together with basic information about them (whether a stem is a noun stem or a verb stem, etc.).
- 2. **Morphotactics:** which classes of morphemes can follow other classes, e.g. the English plural morpheme follows the noun rather than preceding it.
- 3. **Orthographic (spelling) rules** used to model the changes that occur when two morphemes combine, e.g. city + PL = cities.
- A type of lexicon-free FST is **Paice's stemming rules.**

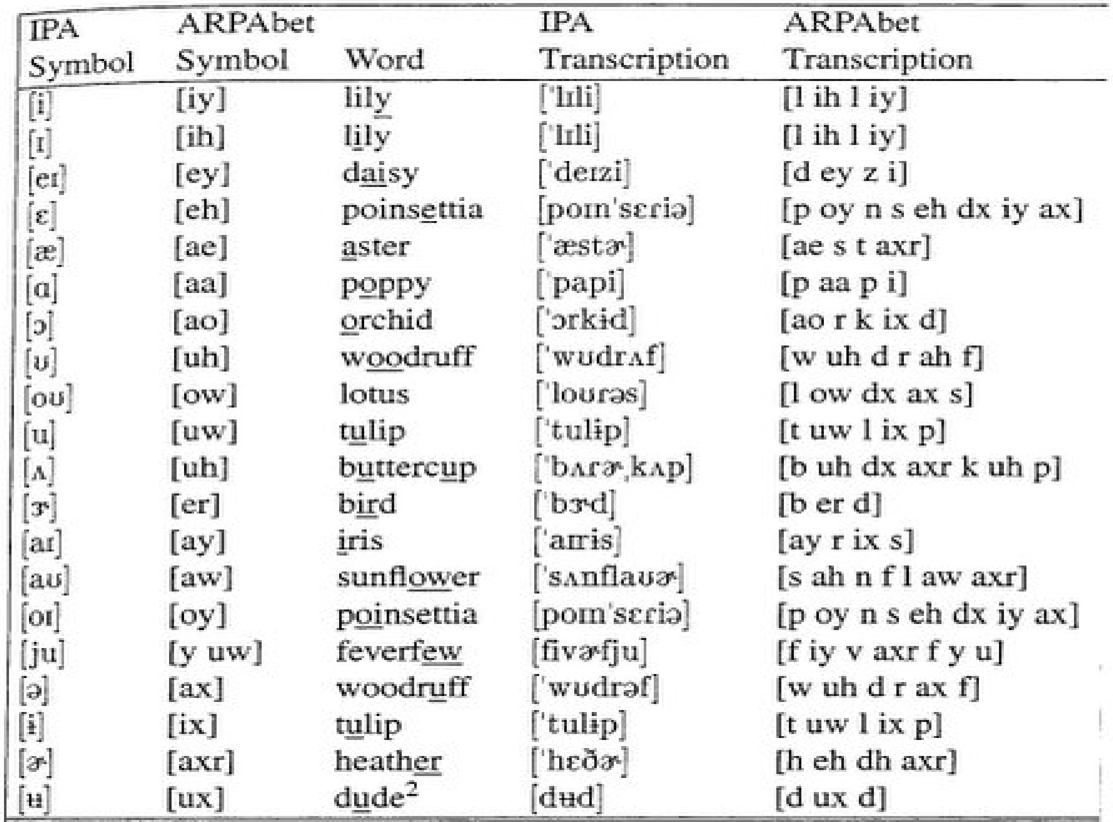
Speech Recognition and Text-to-Speech

- **Phonetics** is the study of the speech sounds used in the languages of the world. Words are pronounced in terms of individual speech units called **phones** or **segments**. There is an **international phonetic alphabet** for describing the pronunciation of any human language. Phonology is the area of linguistics that describes the systematic way that sounds are differently realised in different environments. We describe these different realisations using **phonological rules**. **Phonological learning** is how phonological rules can be automatically induced by machine learning algorithms.



IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[p]	[p]	parsley	['parsli]	[p aa r s l iy]
[t]	[t]	tarragon	['tærəgɑn]	[t ae r ax g aa n]
[k]	[k]	catnip	['kætnip]	[k ae t n ix p]
[b]	[b]	bay	[beɪ]	[b ey]
[d]	[d]	dill	[dɪl]	[d ih l]
[g]	[g]	garlic	['gɑrlɨk]	[g aa r l ix k]
[m]	[m]	mint	[mɪnt]	[m ih n t]
[n]	[n]	nutmeg	['nʌtmɛg]	[n ah t m eh g
[ŋ]	[ng]	ginseng	['dʒɪnsɨŋ]	[jh ih n s ix ng]
[f]	[f]	fennel	['fɛnl̩]	[f eh n el]
[v]	[v]	clove	[kloʊv]	[k l ow v]
[θ]	[th]	thistle	['θɪsl̩]	[th ih s el]
[ð]	[dh]	heather	['hɛðɚ]	[h eh dh axr]
[s]	[s]	sage	[seɪdʒ]	[s ey jh]
[z]	[z]	hazelnut	['heɪzl̩nʌt]	[h ey z el n ah t]
[ʃ]	[sh]	squash	[skwɑʃ]	[s k w a sh]
[ʒ]	[zh]	ambrosia	[æm'broʊʒə]	[ae m b r ow zh ax]
[tʃ]	[ch]	chicory	['tʃɪkɚi]	[ch ih k axr iy]
[dʒ]	[jh]	sage	[seɪdʒ]	[s ey jh]
[l]	[l]	licorice	['lɪkɚɨʃ]	[l ih k axr ix sh]
[w]	[w]	kiwi	['kiwi]	[k iy w iy]
[r]	[r]	parsley	['pɑrsli]	[p aa r s l iy]
[j]	[y]	yew	[yu]	[y uw]
[h]	[h]	horseradish	['hɔrsrædɪʃ]	[h ao r s r ae d ih sh]
[ʔ]	[q]	uh-oh	[ʔʌʔoʊ]	[q ah q ow]
[ɾ]	[dx]	butter	['bʌɾɚ]	[b ah dx axr]
[ɾ̃]	[nx]	wintergreen	[wɪɾ̃ɚgrin]	[w ih nx axr g r i n]
[l̩]	[el]	thistle	['θɪsl̩]	[th ih s el]

Figure 4.1 IPA and ARPAbet symbols for transcription of English consonants.



IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[i]	[iy]	lily	['lɪli]	[l ih l iy]
[ɪ]	[ih]	lily	['lɪli]	[l ih l iy]
[eɪ]	[ey]	daisy	['deɪzi]	[d ey z i]
[ɛ]	[eh]	poinsettia	[poɪn'sɛɾiə]	[p oy n s eh dx iy ax]
[æ]	[ae]	aster	['æstɚ]	[ae s t axr]
[ɑ]	[aa]	poppy	['pɑpi]	[p aa p i]
[ɔ]	[ao]	orchid	['ɔrkɨd]	[ao r k ix d]
[ʊ]	[uh]	woodruff	['wʊdrʌf]	[w uh d r ah f]
[oʊ]	[ow]	lotus	['loʊɾəs]	[l ow dx ax s]
[u]	[uw]	tulip	['tulɨp]	[t uw l ix p]
[ʌ]	[uh]	buttercup	['bʌɾɚˌkʌp]	[b uh dx axr k uh p]
[ɝ]	[er]	bird	['bɝd]	[b er d]
[aɪ]	[ay]	iris	['aɪrɨs]	[ay r ix s]
[aʊ]	[aw]	sunflower	['sʌnflaʊɚ]	[s ah n f l aw axr]
[oɪ]	[oy]	poinsettia	[poɪn'sɛɾiə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	feverfew	[fivɚfju]	[f iy v axr f y u]
[ə]	[ax]	woodruff	['wʊdrəf]	[w uh d r ax f]
[ɨ]	[ix]	tulip	['tulɨp]	[t uw l ix p]
[ɚ]	[axr]	heather	['hɛðɚ]	[h eh dh axr]
[ʉ]	[ux]	dude[2]	[dʉd]	[d ux d]

Figure 4.2 IPA and ARPAbet symbols for transcription of English vowels.

Phonological rules (1)

$$/\mathrm{t}/ \rightarrow [\text{t̪}] \ / \ \underline{\hspace{2em}}\ \theta$$

- Phonological rules: the *t* in *tunafish* is aspirated (followed by a period of voicelessness) while the *t* in *starfish* is not. . Another variant is the dentalised *t* in *eighth.*

Phonological rules (2)

- How do we represent this relation between a *t* and its different realisations in different contexts? We posit an abstract class called the **phoneme**, which is realised as different **allophones** in different contexts. The relationship between a phoneme and its allophones is often captured by writing a phonological rule, e.g. Chomsky and Halle (1968): see below
- Finite state automata are used in various ways to realise phonological rules.

$$/\mathrm{t}/ \rightarrow [\text{t̪}] \;/\; \underline{\qquad\quad}\; \theta$$

Mapping Text to Phones for Text-to-Speech

- An important tool is the **pronunciation dictionary,** e.g. PRONLEX, CMUdict, CELEX (160,000 word forms). CMUdict gives stress levels for the vowels. CELEX gives distinct pronunciations for each part of speech, which helps resolve homographs e.g. *wind (noun)* and *wind (verb)*. However, pronunciation dictionaries tend to be lacking in proper names, and don't deal with things like *Dr. (Doctor? Drive?)*.
- **Beyond dictionary look up: text analysis.**
- Both speech synthesis and speech recognition systems need to be able to guess at the pronunciation of words that are not in their dictionary, e.g. **names**, **morphological productivity** and **numbers**. Medical speech applications such as transcriptions of doctor-patient interviews require pronunciations of names of pharmaceuticals.

An FST-based pronunciation lexicon

- **An FST-based pronunciation lexicon**
- Allen et al. (1987) relied on **letter-to-sound** rules. Each rule specified how a letter or combination of letters was mapped to phones, e.g.
- Fragment [Pronunciation] p-[p]; ph-[f]; phe-[fi]; phes-[fiz]; place-[pleıs]; placi-[pleısi]; plement-[plımεnt];
- Such systems consisted of a long list of such rules and a small list of exceptions. More recent systems have inverted the algorithm, relying on very large dictionaries with letter to sound rules used only for the small number of words that are neither in the dictionary nor are morphological variants (use transducers) of words in the dictionary.

Prosody

- Prosody refers to aspects of pronunciation which cannot be inferred from the sequence of phonemes derived from the dictionary. Three main aspects are **prominence** (stress and accent) **structure** (phrasing, e.g. *I wanted to go to London, but could only get tickets for France)* and **tune** (*Oh really? Oh really!*). Text to speech often sounds wooden, as they aim to produce a neutral declarative version of the input text, spoken in a default way with no reference to discourse history or real world events.

Dealing with Spelling Errors

- spell check on modern word processors
- optical character recognition
- on-line handwriting recognition
- **isolated-word error detection and correction:** correcting spelling errors that result in non-words (e.g. *graffe* for *giraffe*)
- **context-dependent error detection and correction:** using context to detect and correct spelling errors even if they accidentally result in another English word. **Typographical** (e.g. *three* for *there*) or **cognitive** (e.g. *piece* for *peace*)

Minimum edit method of spelling error correction

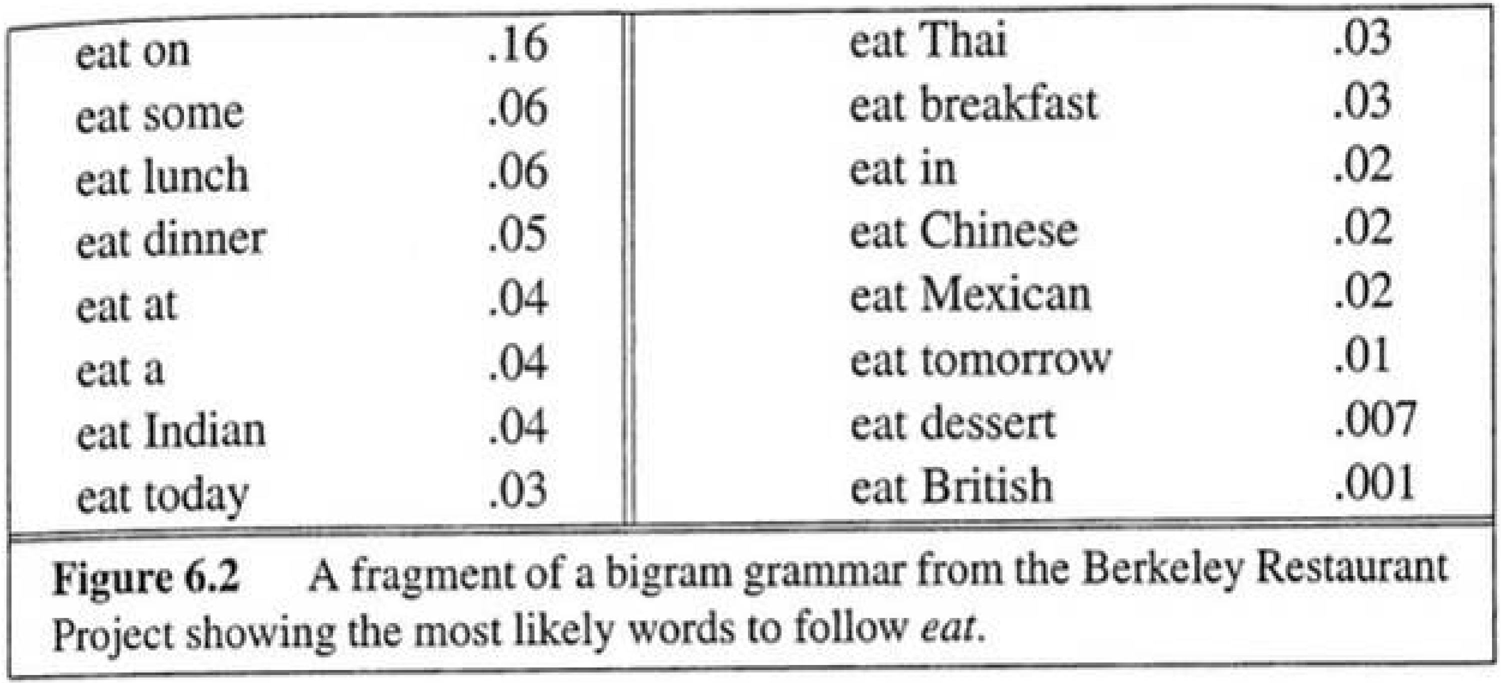
- Damerau (1964) found that 80% of spelling errors in a sample of human keypunched texts were **single-error misspellings**, a single one of the following:
- insertion: mistyping *the* as *ther*
- deletion: mistyping *the* as *th*
- substitution: mistyping *the* as *thw*
- transposition: mistyping *the* as *hte*
- This suggests the **minimum edit** method of spelling error correction. The minimum edits is the least number of insertions, deletions and substitutions required to transform one word into another.
- Exercise: Given a dictionary consisting of *scarf, scare, scene and scent*, what is the most likely correct spelling of *sene?*
- OCR errors are more due to character similarity than keyboard distance, e.g. e/c, m/rn)

Word Prediction and N-Grams

- *I'm going to make a telephone…*
- Word prediction is an essential subtask of speech recognition, augmentative communication for the disabled, context-sensitive spelling error detection, inputting Chinese characters, etc.
- Some attested real-word spelling errors (Kukich, 1992):

- They are leaving in about fifteen minuets.
- The study was conducted be John Black.
- The design an construction of the system will take more than a year.
- Hopefully, all with continue smoothly in my absence.
- He is trying to fine out.

- An N-gram language model uses the previous N-1 words to predict the next one. A bigram is called a first-order Markov Model.
- A fragment of a bigram grammar from the Berkeley Restaurant Project - a speech based restaurant consultant:
- See p 199, and note the formula at the bottom.



eat on	.16	eat Thai	.03
eat some	.06	eat breakfast	.03
eat lunch	.06	eat in	.02
eat dinner	.05	eat Chinese	.02
eat at	.04	eat Mexican	.02
eat a	.04	eat tomorrow	.01
eat Indian	.04	eat dessert	.007
eat today	.03	eat British	.001

Figure 6.2 A fragment of a bigram grammar from the Berkeley Restaurant Project showing the most likely words to follow *eat*.

<s> I	.25	I want	.32	want to	.65	to eat	.26	British food	.60
<s> I'd	.06	I would	.29	want a	.05	to have	.14	British restaurant	.15
<s> Tell	.04	I don't	.08	want some	.04	to spend	.09	British cuisine	.01
<s> I'm	.02	I have	.04	want thai	.01	to be	.02	British lunch	.01

Figure 6.3 More fragments from the bigram grammar from the Berkeley Restaurant Project.

Now we can compute the probability of sentences like *I want to eat British food* or *I want to eat Chinese food* by simply multiplying the appropriate bigram probabilities together, as follows:

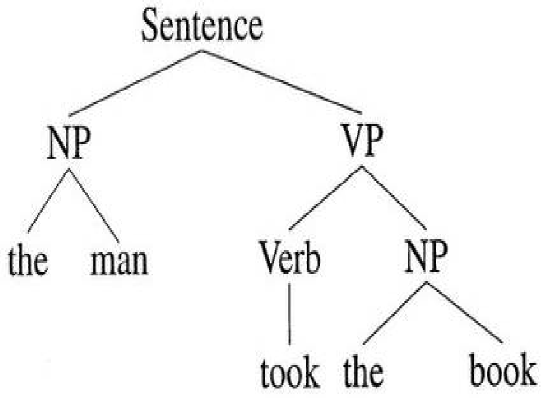
$$\begin{aligned} P(\text{I want to eat British food}) &= P(\text{I}|\text{<s>})P(\text{want}|\text{I})P(\text{to}|\text{want}) \\ &\quad P(\text{eat}|\text{to})P(\text{British}|\text{eat}) \\ &\quad P(\text{food}|\text{British}) \end{aligned}$$

Word Classes and Part-of-Speech Tagging

- No definitive list, but 146 for the C7 tagset (Garside et al., 1997).
- Two broad supercategories: **closed class** and **open class.**
- Main open classes are **nouns** (*cat, Daniel*)**, verbs** (*walk*)**, adjectives** (*green*) and **adverbs** (*slowly*).
- Main closed classes are:
- Prepositions: *on, under, over, near, by, at, from, to, with*
- Determiners: *a, an, the*
- Pronouns: *she, who, I, others*
- Conjunctions: *and, but, or, as, if, when*
- auxiliary verbs: *can, may, should, are*
- particles: *up, down, on, off, in, out, at, by*
- numerals: *one, two, three, first, second, third*
- Tagsets for English, e.g. Penn Treebank
- The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Exercise: manual CLAWS tagger with disambiguation.

Context-free grammars for English

- S → NP + VP
- NP → DET + NOUN
- VP → VERB + NP
- DET → the
- NOUN → man | book
- VERB → took



The first context-free grammar parse tree (Chomsky, 1956)

CFG are also called Phrase-Structure Grammars

- They consist of a set of **rules** or **productions**, each of which expresses the ways that symbols of the language can be grouped together, and a **lexicon** of words or symbols. The symbols that correspond to words in the surface form of the language are called **terminal symbols.**
- The CFG may be thought of in two ways: as a device for generating sentences (**top-down** parsing), or as a device for assigning a structure to a given sentence (**bottom-up** parsing). It is sometimes convenient to represent a **parse-tree** in **bracketed notation** (e.g. the Penn Treebank).
- [S [NP [DET the] [NOUN man]] [VP [VERB took] [NP [DET the] [NOUN book]]]]

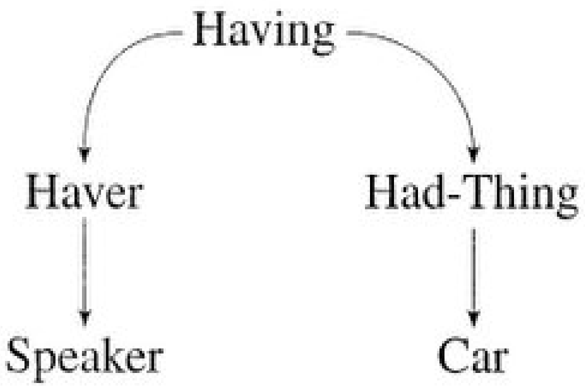
A CFG defines a formal language

- Sentences that can be derived by the grammar are in the formal language defined by that grammar, and are called **grammatical** sentences. Sentences that cannot be derived by a given formal grammar are referred to as **ungrammatical**. In linguistics, the use of formal languages to model natural languages is called **generative grammar**, since the language is defined by the set of possible sentences generated by the grammar.
- Mention **probabilistic grammars**. Choose the most probable interpretation of an ambiguous sentence, by adding weights to the rules.

Semantics

- Semantics is the study of the meaning of linguistic utterances. One way of representing meaning is by using **First Order Predicate Calculus (FOPC),** and we will also look at **Semantic Networks** and **Conceptual Dependency** diagrams.
- See diagrams on p503 (Jurafsky and Martin) and p278 (Rich and Knight).

$$\exists x,y Having(x) \wedge Haver(Speaker,x) \wedge HadThing(y,x) \wedge Car(y)$$



Car
⇑ POSS-BY
Speaker

```
Having
 Haver:   Speaker
 HadThing:  Car
```

Figure 14.1 A list of symbols, two directed graphs, and a record structure: a sampler of meaning representations for *I have a car.*

Computational desiderata for representations

- e.g. Serves (Maharani, VegetarianFood)
- **verifiability** concerns a system’s ability to compare the state of affairs described by a representation to the state of affairs in some world as modelled in a knowledge base.
- e.g. I want to eat somewhere that’s close to the city centre.
- Regardless of any ambiguity in the raw input, it is critical that a meaning representation language support representations that have a single **unambiguous** representation.
- e.g.
- Does Maharani have vegetarian dishes?
- Do they have vegetarian food at Maharani?
- Are vegetarian dishes served at Maharani?
- Does Maharani serve vegetarian fare?
- The notion that inputs that mean the same thing should have the same meaning representation is known as the doctrine of the **canonical form**.

Inference

- e.g Does Maharani serve vegetarian food?
- Serves(Maharani, VegetarianFood) ?
- Serves(x,VegetarianFood) :-
  - VegetarianRestaurant(x).
- VegetarianRestaurant(Maharani).
- Serves(Maharani,VegetarianFood).
- We use the term **inference** to refer to a system's ability to draw valid conclusions based on the meaning representation of inputs and its store of background knowledge.
- Finally, to be useful, a meaning representation scheme must be **expressive** enough to handle an extremely wide range of subject matter.

WordNet (1)

- The **WordNet** thesaurus lists the range of different senses a word can have, and also the range of relations between related word senses:
- hypernym, e.g. breakfast → meal (noun)
- hyponym, e.g. meal → lunch
- has-member e.g. faculty → professor
- member-of e.g. copilot → crew
- has-part e.g. table → leg
- part-of e.g. course → meal

WordNet (2)

- hypernym e.g. fly → travel (verb)
- troponym e.g. walk → stroll
- entails e.g. snore → sleep
- antonym e.g. increase → decrease
- antonym e.g heavy → light (adjective)
- antonym e.g. quickly → slowly (adverb)
- synsets e.g. {chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, soft touch, mug }

The ACAMRIT semantic tagger

- The SEMTAG semantic tagset was originally loosely based on the categories found in the Longman Lexicon of Contemporary English (McArthur, 1981).
- The categories are arranged in a hierarchy, with 21 major discourse fields denoted by an upper case letter (such as E for "emotional actions, states and processes"), then divided and sometimes even subdivided again.
- This is shown using numeric components of the semantic codes such as 4.1.
- Antonyms are identified using the symbols + and -. Thus "happy" is normally tagged E4.1+, and "sad" is normally tagged as E4.1-. Comparatives are shown with ++ or --, and superlatives with +++ or ---.

ACAMRIT (2)

- In some cases, a word type can only represent one possible category. Often, however, a word such as “spring” can have a number of different meanings, each requiring a different semantic tag. In such cases, disambiguation is achieved using six types of additional evidence, as follows:

Additional Evidence for WSD

- The POS tag assigned by CLAWS. For example, if “spring” is a verb, we know it must mean “jump”.
- The general likelihood of a word taking a particular meaning, as found in certain frequency dictionaries.
- Idiom lists are kept. If an entire idiomatic phrase is found in the text being analysed, it is assumed that the idiomatic meaning of each word in the phrase is more likely than individual interpretations of the words.
- The domain of discourse can be an indicator. For example, if the topic of discussion is footwear, then “boot” is unlikely to refer to the boot of a car.
- Special rules have been developed for the auxiliary verbs “be” and “have”.
- Proximity disambiguation. Are any collocates of the word, suggesting a particular interpretation, found in the immediate vicinity?

Tagset

- The full set of SEMTAG semantic tags can be found on http://www.comp.lancs.ac.uk/ucrel/acamrit/setags.txt.

Some ACAMRIT codes

- G1 Government, Politics and elections
- G1.1 Government etc.
- G1.2 Politics
- G2 Crime, law and order
- G2.1 Crime, law and order: Law and order
- G2.2 General ethics
- G3 Warfare, defence and the army; weapons
- H1 Architecture and kinds of houses and buildings
- H2 Parts of buildings
- H3 Areas around or near houses
- H4 Residence
- H5 Furniture and household fittings
- I1 Money generally
- I1.1 Money: Affluence
- I1.2 Money: Debts
- I1.3 Money: Price
- I2 Business
- I2.1 Business: Generally
- I2.2 Business: Selling
- I3 Work and employment
- I3.1 Work and employment: Generally
- I3.2 Work and employment: Professionalism
- I4 Industry

Text tagged with Part Of Speech and Semantic Code

I_PPIS1_Z8mf went_VVD_M1[i3.2.1 down_RP_M1[i3.2.2
yesterday_RT_T1.1.1 to_II_Z5 the_AT_Z5 Peiraeus_NP1_Z99
with_IW_Z5 Glaucon_NP1_Z99 ,_,_PUNC the_AT_Z5
son_NN1_S4m of_IO_Z5 Ariston_NP1_Z99 ,_,_PUNC
to_TO_Z5 pay_VVI_I1.2 my_APPGE_Z8 devotions_NN2_Z99
to_II_Z5 the_AT_Z5 Goddess_NN1_S9/S2.1f ,_,_PUNC
and_CC_Z5 also_RR_N5++ because_CS_Z5 I_PPIS1_Z8mf
wished_VVD_X7+ to_TO_Z5 see_VVI_X3.4 how_RRQ_Z5
they_PPHS2_Z8mfn would_VM_A7+ conduct_VVI_A1.1.1
the_AT_Z5 festival_NN1_K1/S1.1.3+ since_CS_Z5
this_DD1_Z8 was_VBDZ_A3+ its_APPGE_Z8
inauguration_NN1_Z99 ._._PUNC

Discourse

- Up to now, we have focussed mainly on language pheomena that operate at the word or sentence level. Of course, language does not normally consist of isolated, unrelated sentences, but instead of related groups of sentences. We refer to such a group of sentences as a discourse.
- Coherence and reference are discourse phenomena: consider
- *John went to Bill's car dealership to check out an Acura Integra. He looked at it for about an hour.*
- Automatic reference resolution depends mainly on proximity rules and constraints on coreference, e.g. agreement in gender, number and animacy.

Discoursal Annotation

- (0) The state Supreme Court has refused to release {1[2 Rahway State Prison 2] inmate 1} (1 James Scott 1) on bail. (1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction. (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision. Meanwhile {3 <1 his promoter 3}, { 3 Murad Mohammad 3}, said Wednesday <3 he netted only $15,250 for (4 [1 Scott 1] ‘s nationally televised fight against {5 ranking contender 5 } (5 Yaqui Lopez 5) last Saturday 4).

Dialogue Acts (Bunt) or Conversational Moves (Power)

- **STATEMENT A claim made by the speaker**
- **INFO-REQUEST A question by the speaker**
- CHECK A question for confirming information
- **INFLUENCE-ON-ADDRESSEE (= Searle's directives)**
- OPEN-OPTION A weak suggestion or listing of options
- ACTION-DIRECTIVE An actual command
- **INFLUENCE-ON-SPEAKER(= Austin's commissives)**
- OFFER Speaker offers to do something (subject to confirmation)
- COMMIT Speaker is committed to doing something
- **CONVENTIONAL Other**
- OPENING Greetings
- CLOSING Farewells
- THANKING Thanking and responding to thanks.

Machine Translation

- Rough translations can be post-edited by a human translator, which may speed up the translation process. This is **computer-aided human translation** (CAHT) rather than fully-automated machine translation.
- Weather forecasting is an example of a **sublanguage** domain that can be modelled completely enough to use raw MT output even without postediting. The domain has a limited vocabulary and only a few basic phrase types. Ambiguity is rare. Other domains that are sub-language like include equipment maintenance manuals, air travel queries, appointment scheduling, restaurant recommendations, hotel bookings.
- {see diagram p. 806 - difficulty even in single word translation}
- {see diagram p. 814 - the relation between the transfer and interlingua models}

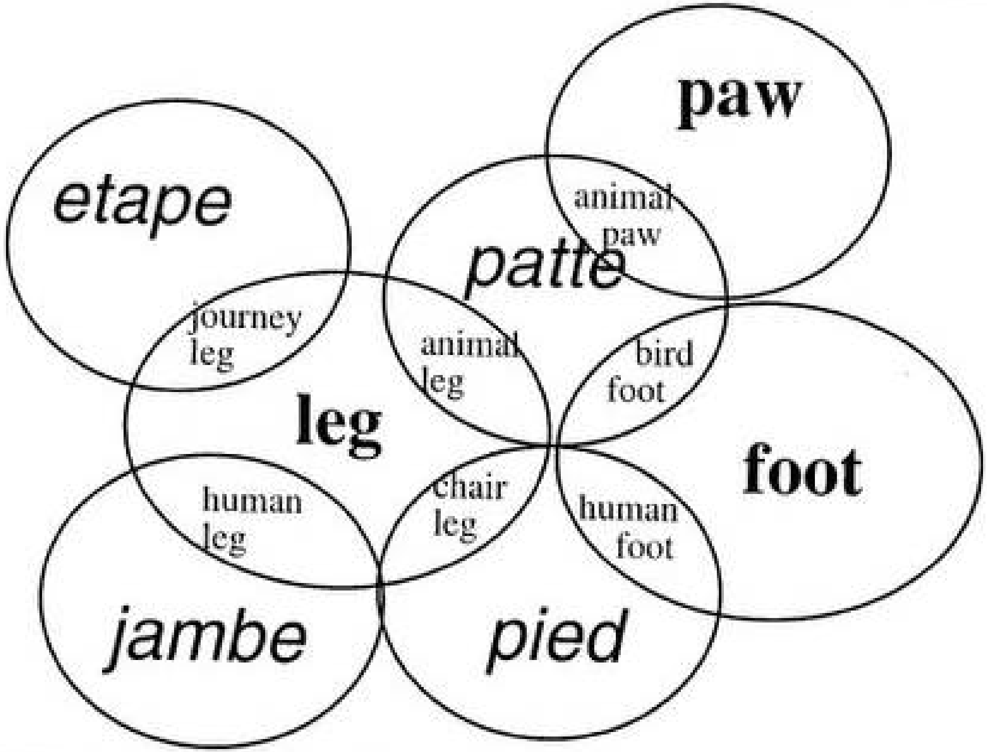


Figure 21.2 The complex overlap between English *leg*, *foot*, etc, and various French translations.

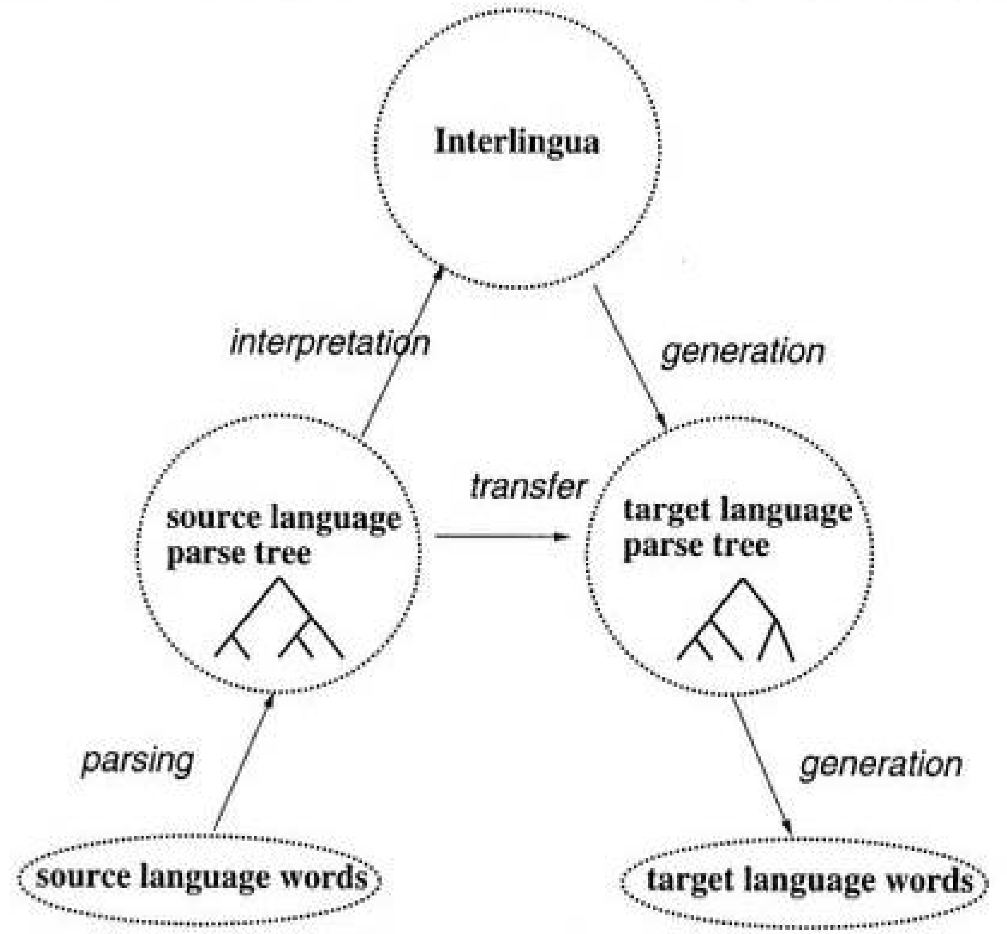


Figure 21.7 Diagram Suggesting the Relation Between the Transfer and Interlingua Models, generally credited to Vauqois.

Using statistical techniques

- Human translators aim for a balance between **faithfulness** and **fluency.**
- Best-translation T = argmax fluency(T) faithfulness (T,S)
- where T is the target language sentence and S is the source language sentence. This model of translation was first described by researchers coming from speech recognition (Brown et al., 1990, 1993) and has clear parallels with Bayesian reasoning.
- Best-translation T = argmax P(T) P(S|T)
- P(T) can be estimated using **n-gram models**
- P(S|T) is the product of the probabilities that each target language word is a possible translation of some source language word. For this we need to know, for every source language word, the probability of it mapping to each possible target language word.

Probabilistic Dictionaries

- Standard bilingual dictionaries do not have this information, but we can derive **probabilistic dictionaries** from **aligned corpora**. (Sentence alignment, followed by the more difficult task of Word alignment (Melamed)).
- Example of a probabilistic dictionary (Brown et al. 1990) for the English word *not*.
- **French Probability** pas 0.469; ne 0.460; non 0.024; pas du tout 0.003; faux 0.003; plus 0.002; ce 0.002; que 0.002; jamais 0.002;
- **FertilityProbability** 2 0.758; 0 0.133; 1 0.106;
- **Example-based machine translation** and **translation memory** and neural translation.