Language and Computers (Ling 384)	Language and Computers Topic 4: Writer's aids Introduction Error causes Kostead embylings Ponetic emrs Kroatedge problems Difficult issues	Who cares about spelling? Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the Itteers in a wrod are, the olny iprmoetnt tihng is taht the frist and Isat Itteer be at the rghit pclae. The rset	Language and Computers Topic 4: Writer's aids Introduction Error causes Koekedge problems Nowledge problems Difficult issues	 Why people care about spelling Misspellings can cause misunderstandings and real-life problems: For example: 	Language and Computers Topic 4: Writer's aids Introduction Error causes Koyboard mitspings Phonetic errors Knowledge problems Difficult issues
Topic 4: Writer's aids (Spelling and Grammar Correction) Detmar Meurers* Dept. of Linguistics, OSU	Difficult results Tokenization Infection Productivity Non-word error detection Dictionaries N-gram analysis Isolated-word error correction	can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey Iteter by istlef, but the wrod as a wlohe. (See http://www.mrc-cbu.cam.ac.uk/personal/matt.davis/Cmabrigde/ for	Diminizini Issues Tolenization Inflection Productivity Non-word error detection Dictionaries N-gram analysis Isolated-word error correction	 > Did you see her god yesterday? It's a big golden retriever. > This will be a fee [free] concert. > 1991 Bell Atlantic & Pacific Bell telephone network outages were partly caused by a typographical error: A 6 in a line of computer code was supposed to be a D. 	Dimicult issues Tokenization Inflection Productivity Non-word error detection Dictionaries N-gram analysis Isolated-word erro correction
Autumn 2006	Correction Rule-based methods Similarly key techniques Probabilistic methods Minimum edit distance Grammar correction Syntax Computing with Syntax Grammar correction rules	the story behind this supposed research report.) A dtcoor has aimttded the magItheuansr of a taegene cceanr ptinaet who deid aetfr a haptosil durg bednlur.	COTRECTION Rule-based methods Similarity key techniques Probabilistic methods Minimum edit distance Grammar correction Syntax Computing with Syntax Grammar correction rules	"That one error caused the equipment and software to fail under an avalanche of computer-generated messages." (Wall Street Journal, Nov. 25, 1991)	COTRECTION Rule-based methods Similarity key techniques Probabilitatic methods Minimum edit distance Grammar correction Syntax Computing with Syntax Grammar correction rules
	Caveat emptor		Caveat emptor		Caveat emptor
Why people care about spelling (cont.)	Language and Computers Topic 4: Writer's aids	How are spell checkers used?	Language and Computers Topic 4: Writer's aids	Detection vs. Correction	Language and Computers Topic 4: Writer's aids
 Standard spelling makes it easy to organize words and text: e.g., Without standard spelling, how would you look up things in a lexicon or thesaurus? e.g., Optical character recognition software can use knowledge about standard spelling to recognize scanned words even for hardly legible input. Standard spelling makes it possible to provide a single text, which is accessible to a wide range of readers (different backgrounds, speaking different dialects, etc.). Using standard spelling is associated with being well-educated, i.e., is used to make a good impression in social interaction. 	Introduction Intro	 interactive spelling checkers = spell checker detects errors as you type. It may or may not make suggestions for correction. Requires a "real-time" response (i.e., must be fast) It is up to the human to decide if the spell checker is right or wrong. If there are a list of choices, we may not require 100% accuracy in the corrected word automatic spelling correctors = spell checker runs on a whole document, finds errors, and corrects them A much more difficult task. A human may or may not proofread the results later. 	Introduction In	 There are two distinct tasks: error detection = simply find the misspelled words error correction = correct the misspelled words e.g., It might be easy to tell that <i>ater</i> is a misspelled word, but what is the correct word? <i>water</i>? <i>later</i>? <i>after</i>? ⇒ Depends on what we want to do with our results as to what we want to do. Note, though, that detection is a prerequisite for correction. 	Introduction Free Causes Maybeart manyings Maybeart manyings Maybeart Marken and Marken and Mar
What causes errors?	Language and Computers Topic 4: Writer's aids Introduction Error causes Keyboard mispings	Keyboard mistypings	Language and Computers Topic 4: Writer's aids Introduction Error causes Keyboard misipings	Keyboard mistypings (cont.) Keyboard proximity	Language and Computers Topic 4: Writer's aids Introduction Error causes Keyboard mispings
 Keyboard mistypings Phonetic errors Knowledge problems 	Productie entry of Koneledge problems Difficult issues Difficult issues Persoarchwig Non-word error detection Discoarsies N-gram analysis Isolated-word error correction Rust-based methods Brindlarly key techniques Probabiliste methods Meinrum edit distance Gramma correction nues Capacity and Syntas Campuing with Syntas	 Space bar issues run-on errors = two separate words become one e.g., the fuzz becomes thefuzz split errors = one word becomes two separate items e.g., equalization becomes equali zation Note that the resulting items might still be words! e.g., a tollway becomes atoll way 	Resetterens Konskräge problem Difficult issues Difficult issues Resetution Productivity Non-word error detection Dictionaries Nigram analysis Isolated-word error correction Rule based methods Similarity kay techniques Probabiliss methods Meimum edit distance Grammar correction Synax Caputary dith Syntas Grammar correction nules	 e.g., <i>Jack</i> becomes <i>Hack</i> since <i>h</i> and <i>j</i> are next to each other on a typical American keyboard Physical similarity similarity of shape, e.g., mistaking two physically similar letters when typing up something handwritten e.g., <i>tight</i> for <i>fight</i> 	Producti series Nonsierdige problems Difficult issues Difficult issues Infector Productive Areas Non-World error detection Distantes Norward error detection Sisolated-Word error correction Rule based rethols Similarity kay tothiques Probabilicar features Meinum est distance Grammar correctiol Synax Computer guith Synta Grammer correction rules Caveat emptor

Phonetic errors	Language and Computers Topic 4: Writer's aids	Phonetic errors (cont.)	Language and Computers Topic 4: Writer's aids	More examples for phonetic errors	Language and Computers Topic 4: Writer's aids
 phonetic errors = errors based on the sounds of a language (not necessarily on the letters) homophones = two words which sound the same e.g., red/read (past tense), cite/site/sight, they're/their/there Spoonerisms = switching two letters/sounds around e.g., It's a tavy grain with biscuit wheels. 	Write's aids Write's aids Error causes Kaytead mayings Prometic enco Monower and an anti- Difficult issues Difficult issues Difficult issues Difficult issues Difficult issues Non-word error detection Non-word error detection Similar to the	 letter substitution: replacing a letter (or sequence of letters) with a similar-sounding one e.g., John kracked his nuckles. instead of John cracked his knuckles. e.g., I study sikologee. word replacement: replacing one word with some similar-sounding word e.g., John battled me on the back. instead of John patted me on the back. 	Write's alos Write's alos Error causes Kaytead maying Prantisent Kowledge problem Difficult issues Difficult issues Difficult issues Difficult issues Difficult issues Production Non-word error detection Non-word error correction Nate has dembnds Similarly kay techniques Probabilistic methods Minimum edi distance Grammar correction Syntax Computing with Syntax Computing with Syntax Computing with Syntax	 (1) a. death in Venice b. deaf in Venice (2) a. give them an ice bucket b. give them a nice bucket (3) a. the stuffy nose b. the stuff he knows (4) a. the biggest hurdle b. the biggest turtle (5) a. some others b. some mothers (6) a. a Coke and a danish b. a coconut danish 	Writer's adds Introduction Error causes Naybaud matepings Production Difficult issues Difficult issues Diffi
Knowledge problems	10/72 Language and Computers Topic 4: Writer's aids	What makes spelling correction difficult?	11/72 Language and Computers Topic 4: Writer's aids	Tokenization	12/72 Language and Computers Topic 4: Writer's aids
 not knowing a word and guessing its spelling (can be phonetic) e.g., sientist not knowing a rule and guessing it e.g., Do we double a consonant for ing words? jog → joging joke → jokking 	Inter Salos Intervention Error Causes Maybeard mathyning Partial Martingen Marting	 Tokenization: What is a word? Inflection: How are some words related? Productivity of language: How many words are there? How we handle these issues determines how we build a dictionary. 	Intent saus Introduction Error causes Moybaard misping Posadic errors Norwedge problem Difficult issues Difficult issues Mon-word error detection Non-word error detection Non-word	 spaces, but this is not always so clear. contractions = two words combined into one e.g., can't, he's, John's [car] (vs. his car) multi-token words = (arguably) a single word with a space in it e.g., New York, in spite of, deja vu hyphens (note: can be ambiguous if a hyphen ends a line) Some are always a single word: e-mail, co-operate Others are two words combined into one: Columbus-based, sound-change Abbreviations: may stand for multiple words e.g., etc. = et cetera, ATM = Automated Teller Machine 	Introduction Introduction Encoduction May an encoduction May an encodu
 Inflection A word in English may appear in various guises due to word inflections = word endings which are fairly systematic for a given part of speech plural noun ending: the boy + s → the boys past tense verb ending: walk + ed → walked This can make spell-checking hard: There are exceptions to the rules: mans, runned There are words which look like they have a given ending, but they don't: Hans, deed 	13/72 Language and Computers Topic 4: Writer's aids Introduction Error causes Kaybear mispings Prioratic errors Anovelege problems Difficult issues Difficult issues Difficult issues Difficult issues Difficult issues Non-word error detection Non-word error detection Rule based methods Similarity bay techniques Probabilistic methods Margum set distance Grammar correction Syntae Compacting with Syntas Compacting Co	 Productivity part of speech change: nouns can be verbified emailed is a common new verb coined after the noun email morphological productivity: prefixes and suffixes can be added e.g., I can speak of un-email-able for someone who you can't reach by email. words entering and exiting the lexicon, e.g.: thou, or spleet 'split' (Hamlet III.2.10) are on their way out d'oh seems to be entering 	14/72 Language and Computers Topic 4: Writer's aids Introduction Error causes Kaybear mispings Pionatic errors Kaybear mispings Pionatic errors Difficult issues Difficult issues Difficult issues Non-word error detection Non-word error detection Rule based mehods Similary say techniques Probabilist mehods Marxim edi diatares Grammar correction Nets Comparing with Spritzs Comparing	 Techniques used for spell checking Non-word error detection Isolated-word error correction Context-dependent word error detection and correction → grammar correction. 	15/72 Language and
	Caveat emptor 16/72		Caveat emptor 17/72		Caveat emptor 18/72

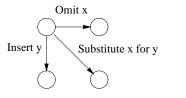
Non-word error detection	Language and Computers Topic 4:	Dictionaries	Language and Computers Topic 4:	Dictionary construction	Language and Computers Topic 4:
 non-word error detection is essentially the same thing as word recognition = splitting up "words" into true words and non-words. How is non-word error detection done? using a dictionary (construction and lookup) n-gram analysis 	Liple 4: Write's aids Introduction Error causes Reposed mispings Phonetic error Contention Internation	 Intuition: Have a complete list of words and check the input words against this list. If it's not in the dictionary, it's not a word. Two aspects: Dictionary construction = build the dictionary (what do you put in it?) Dictionary lookup = lookup a potential word in the dictionary (how do you do this quickly?) 	Hufe-aids Hitroduction Error causes Kookaad mispings Phoetic enrors Kookaad mispings Difficult issues Difficult issues Difficult issues Non-word error deficition Non-word error deficition Non-word error deficition Non-word error deficition Non-word error deficition Singleicher Ngmananjsis Sicolated-word error correction Singleicher Phadelister mehods Singleicher Phadelister mehods Singleicher Phadelister mehods Singleicher Correction Singleicher Correction Singleicher Computing with Syntas Grammar correction was Caveat emptor	 Do we include inflected words? i.e., words with prefixes and suffixes already attached. Pro: lookup can be faster Con: takes much more space, doesn't account for new formations (e.g., <i>google → googled</i>) Want the dictionary to have only the word relevant for the user → domain-specificity e.g., For most people <i>memoize</i> is a misspelled word, but in computer science this is a technical term and spelled correctly. Foreign words, hyphenations, derived words, proper nouns, and new words will always be problems for dictionaries since we cannot predict these words until humans have made them words. Dictionary should probably be dialectally consistent. e.g., include only <i>color</i> or <i>colour</i> but not both 	Writer's aids Introduction Error causes Medical mappings Phonaits enror Medical mappings Phonaither Internation Difficult issues Phonaither Pho
Dictionary lookup	Language and Computers Topic 4: Writer's aids	N-gram analysis	20/72 Language and Computers Topic 4: Writer's aids	How do we store and use n-gram information?	21/72 Language and Computers Topic 4: Writer's aids
 Several issues arise when trying to look up a word: Have to make lookup fast by using efficient lookup techniques, such as a hash table. Have to strip off prefixes and suffixes if the word isn't an entry by itself. <i>running → run</i> <i>nonreligiously → religious</i> 	<section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	 An n-gram here is a string of <i>n</i> letters. a 1-gram (unigram) at 2-gram (bigram) ate 3-gram (trigram) late 4-gram : : We can use this n-gram information to define what the possible strings in a language are. e.g., po is a possible English string, whereas kvt is not. 	<section-header><section-header><section-header><text><text><text><text><text></text></text></text></text></text></section-header></section-header></section-header>	 Language Identification). But, maybe we just want to know if an n-gram is possible. We could have a list of possible and impossible n-grams (1 = possible, 0 = impossible): po 1 kvt 0 police 1 asdf Any word which has a 0 for any substring is a misspelled word. Problems with such an approach: Information is repeated (<i>po</i> is in <i>police</i>) Requires a lot of computer storage space Inefficient (slow) when looking up every string 	Introduction Encauses Mayner measures Mayner measures
 Bigram array Instead, we can define a bigram array = information stored in a tabular fashion. An example, for the letters k, l, m, with examples in parentheses k n t k n t (elk) t (hello) t (alms) n 0 t (hammer) i The first letter of the bigram is given by the vertical letters (i.e., down the side), the second by the horizontal ones (i.e., across the top). This is a non-positional bigram array = the array 1's and 0's apply for a string found anywhere within a word (beginning, 4th character, ending, etc.). 	Larrie Language and Computers Topic 4: Writer's aids Introduction Error causes Notwedge problems Difficult issues Notwedge problems Difficult issues Difficult issues Difficu	 Positional bigram array To store information specific to the beginning, the end, or some other position in a word, we can use a positional bigram array = the array only applies for a given position in a word. Here's the same array as before, but now only applied to word endings: i.i. k 	Lasrie Language and Longues Topic 4: Writer's aids Introduction Erroduction Erroduction Biotection Difficult issues Difficult is	 Isolated-word error correction Having discussed how errors can be detected, we want to know how to correct these misspelled words: The most common method is isolated-word error correction = correcting words without taking context into account. Note: This technique can only handle errors that result in non-words. Knowledge about what is a typical error helps in finding correct word. 	Language and Computers Topic 4: Writer's aids Introduction Error Causes Noted spectrosmology Noted Spectrosmology

				1	
Knowledge about typical errors	Language and Computers Topic 4: Writer's aids	Isolated-word error correction methods	Language and Computers Topic 4: Writer's aids	Rule-based methods	Language and Computers Topic 4: Writer's aids
	Introduction	Many different methods are used; we will briefly look at	Introduction	One can generate correct spellings by writing rules:	Introduction
	Error causes	four methods:	Error causes		Error causes
 word length effects: most misspellings are within two 	Keyboard mistypings Phonetic errors	► rule-based methods	Keyboard mistypings Phonetic errors	 Common misspelling rewritten as correct word: 	Keyboard mistypings Phonetic errors
characters in length of original	Knowledge problems	 similarity key techniques 	Knowledge problems Difficult issues	► e.g., $hte \rightarrow the$	Knowledge problems
\rightarrow When searching for the correct spelling, we do not	Tokenization	 minimum edit distance 	Tokenization	 Rules based on inflections: 	Tokenization
usually need to look at words with greater length differences.	Inflection Productivity	 probabilistic methods 	Productivity	• based on inflections. • e.g., VCing \rightarrow VCCing, where	Productivity
	Non-word error detection	The methods play a role in one of the three basic steps:	Non-word error detection	V = letter representing vowel,	Non-word error detection
first-position error effects: the first letter of a word is	Dictionaries N-gram analysis	1. Detection of an error (discussed above)	Dictionaries N-gram analysis	basically the regular expression [aeiou]	Dictionaries N-gram analysis
rarely erroneous	Isolated-word error	2. Generation of candidate corrections	Isolated-word error	C = letter representing consonant, basically [bcdfghjklmnpqrstvwxyz]	Isolated-word error correction
\rightarrow When searching for the correct spelling, the process is	Correction Rule-based methods Similarity key techniques	 rule-based methods 	Correction Rule-based methods Similarity key techniques	 based on other common spelling errors (such as 	Rule-based methods Similarity key techniques
sped up by being able to look only at words with the same first letter.	Probabilistic methods Minimum edit distance	 similarity key techniques Depling of condidate corrections 	Probabilistic methods Minimum edit distance	keyboard effects or common transpositions):	Probabilistic methods Minimum edit distance
	Grammar correction	 3. Ranking of candidate corrections probabilistic methods 	Grammar correction	► e.g., $CsC \rightarrow CaC$	Grammar correction
	Syntax Computing with Syntax	 probabilistic methods minimum edit distance 	Syntax Computing with Syntax	► e.g., Cie → Cei	Syntax Computing with Syntax
	Grammar correction rules Caveat emptor		Grammar correction rules		Grammar correction rules
			Cavear emptor		
	28/72 Language and		29/72		30/72 Language and
Similarity key techniques	Computers Topic 4:	Probabilistic methods	Language and Computers Topic 4:	Confusion probabilities	Computers Topic 4:
	Writer's aids		Writer's aids	 For the various reasons discussed above (keyboard 	Writer's aids
Problem: How can we find a list of possible corrections?	Introduction	Two main probabilities are taken into account:	Introduction	layout, phonetic similarity, etc.) people type other letters than the ones they intended.	Introduction
Solution: Store words in different boxes in a way that	Error causes Keyboard mistypings		Error causes Keyboard mistypings	 It is impossible to fully investigate all possible error 	Error causes Keyboard mistypings
puts the similar words together.	Phonetic errors Knowledge problems	transition probabilities = probability (chance) of going	Phonetic errors Knowledge problems	causes and how they interact, but we can learn from	Phonetic errors Knowledge problems
Example:	Difficult issues	from one letter to the next.	Difficult issues	watching how often people make errors and where.	Difficult issues
 Start by storing words by their first letter (first letter official) 	Tokenization	e.g., What is the chance that a will follow p in English? That would follow p?	Tokenization Inflection	One way of doing so is to build a confusion matrix = a	Tokenization
effect), e.g., <i>punc</i> starts with the code P.	Productivity Non-word error	That <i>u</i> will follow <i>q</i> ?	Productivity Non-word error	table indicating how often one letter is mistyped for	Productivity Non-word error
2. Then assign numbers to each letter	detection Dictionaries	confusion probabilities = probability of one letter	detection Dictionaries	another	detection Dictionaries
 e.g., 0 for vowels, 1 for b, p, f, v (all bilabials), and so 	N-gram analysis	being mistaken (substituted) for another (can be derived	N-gram analysis	correct	N-gram analysis
forth, e.g., punc \rightarrow P052	Isolated-word error correction	from a confusion matrix)	Isolated-word error correction	r s t	Isolated-word error correction
3. Then throw out all zeros and repeated letters,	Rule-based methods Similarity key techniques Probabilistic methods	e.g., What is the chance that q is confused with p?	Rule-based methods Similarity key techniques Probabilistic methods		Rule-based methods Similarity key techniques
▶ e.g., P052 → P52.	Minimum edit distance	Useful to combine probabilistic techniques with dictionary	Minimum edit distance	r n/a 12 22	Minimum edit distance
Look for real words within the same box,	Grammar correction Syntax	methods	Grammar correction Syntax	typed s 14 n/a 15	Grammar correction Syntax
 e.g., <i>punk</i> is also in the P52 box. 	Computing with Syntax Grammar correction rules		Computing with Syntax Grammar correction rules	t 11 37 n/a	Computing with Syntax Grammar correction rules
	Caveat emptor		Caveat emptor		Caveat emptor
	31/72		32/72	(cf. Kernighan et al 1999)	33/72
How is a mistyped word related to the intended?	Language and	Minimum edit distance	Language and	Computing edit distances	Language and
now is a mistyped word related to the interface.	Computers Topic 4:		Computers Topic 4:	Figuring out the worst case	Computers Topic 4:
Types of operations	Writer's aids		Writer's aids		Writer's aids
	Introduction	. In order to real appoints applied continue time, it can be	Introduction	To be able to compute the edit distance of two words at	Introduction
insertion = a letter is added to a word	Error causes Keyboard mistypings	 In order to rank possible spelling corrections, it can be useful to calculate the minimum edit distance = 	Error causes Keyboard mistypings	all, we need to ensure there is a finite number of steps.	Error causes Keyboard mistypings
deletion = a letter is deleted from a word	Phonetic errors Knowledge problems	minimum number of operations it would take to convert	Phonetic errors Knowledge problems	This can be accomplished by	Phonetic errors Knowledge problems
substitution = a letter is put in place of another one	Difficult issues	one word into another.	Difficult issues	 requiring that letters cannot be changed back and forth 	Difficult issues
transposition = two adjacent letters are switched	Inflection Productivity	 For example, we can take the following five steps to 	Inflection Productivity	a potentially infinite number of times, i.e., we	Inflection Productivity
Note that the first two alter the length of the word, whereas	Non-word error	convert junk to haiku:	Non-word error	 limit the number of changes to the size of the material we are presented with, the two words. 	Non-word error
the second two maintain the same length.	detection Dictionaries	1. $junk \rightarrow juk$ (deletion)	detection Dictionaries	 Idea: Never deal with a character in either word more 	detection Dictionaries
	N-gram analysis Isolated-word error	2. $juk \rightarrow huk$ (substitution)	N-gram analysis Isolated-word error	than once.	N-gramanalysis
General properties	Correction Rule-based methods	3. $huk \rightarrow hku$ (transposition) 4. $hku \rightarrow hiku$ (insertion)	Correction Rule-based methods	► Result:	Correction Rule-based methods
single-error misspellings = only one instance of an	Similarity key techniques Probabilistic methods	5. $hiku \rightarrow haiku$ (insertion)	Similarity key techniques Probabilistic methods	 In the worst case, we delete each character in the first 	Similarity key techniques Probabilistic methods
error	Minimum edit distance	 But is this the minimal number of steps needed? 	Minimum edit distance	word and then insert each character of the second	Minimum edit distance
multi-error misspellings = multiple instances of errors	Grammar correction Syntax		Grammar correction Syntax	word.	Grammar correction Syntax
(harder to identify)	Computing with Syntax Grammar correction rules		Computing with Syntax Grammar correction rules	 The worst case edit distance for two words is length(word1) + length(word2) 	Computing with Syntax Grammar correction rules
	Caveat emptor		Caveat emptor		Caveat emptor
	34/72		35/72		36/72

Computing edit distances

Using a graph to map out the options

- To calculate minimum edit distance, we set up a directed, acyclic graph, a set of nodes (circles) and arcs (arrows).
- Horizontal arcs correspond to deletions, vertical arcs correspond to insertions, and diagonal arcs correspond to substitutions (and a letter can be "substituted" for itself).



Language and Computing edit distances An example graph

Topic 4

Writer's aids

Introduction

Error causes

Keyboard mistyp

Difficult issues

Non-word error

Isolated-word error

detection

N-gram analysi

correction

Rule-bared m

larity key te

Grammar correction

Computing with Synta

Caveat emptor

Language and

Topic 4:

Writer's aids

Introduction

Error causes

Difficult issues

Non-word error

Isolated-word erro

arity key techni

Grammar correction

Computing with Suptr

Caveat emptor

Language and

Computers

Topic 4:

Writer's aids

Introduction

Error causes

Phonetic errors

Difficult issues

Non-word error

Isolated-word erro

Similarity key techniqu

Minimum edit distance

nputing with Synta

Grammar cor

Caveat empto

Inflection

Productivity

detection

N-gram analysis

correction

Probabilie

Rule-based met

40/7

Productivity

detection

N-gram analysi

Say, the user types in plog.

Computing edit distances

with the least cost.

How to compute the path with the least cost

paths to check.

- We want to calculate how far away peg is (one of the possible corrections). In other words, we want to calculate the minimum edit distance (or minimum edit cost) from plog to peg.
- As the first step, we draw the following directed graph:

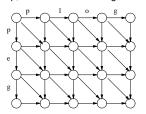
We want to find the path from the start (1) to the end (20)

how many changes we have to make.

Follow every path from start (1) to finish (20) and see

But this is very inefficient! There are many different

The simple but dumb way of doing it:



Language and Computing edit distances Adding numbers to the example graph Writer's aids

Topic 4

Error causes

Keyboard misty

Difficult issues

Non-word error

Isolated-word erro

Tokenizati

detection

N-gram analys

correction

Rule-based

Proba

larity key te

Grammar correcti

Computing with Synta

Caveat emptor

Phonetic errors

Difficult issues

Non-word error

Isolated-word erro

Similarity key techniqu

Minimum edit distance

Computing with Synta

Caveat emptor

Grammar cor

Tokenizati

Inflection

Productivity

detection

N-gram analysi

correction

Probabil

Rule-based met

Dictic

Error causes The graph is acvclic = for any given node, it is impossible to return to that node by following the arcs. We can add identifiers to the states, which allows us to define a topological order: detection Isolated-word erro correction Pula-barada larity key t Grammar correction

Language and Language and Computing edit distances Topic 4: The smart way to compute the least cost Writer's aids Introduction Introduction Error causes ► The smart way to compute the least cost uses dynamic Error causes programming = a program designed to make use of results computed earlier Difficult issues Difficult issues Toks We follow the topological ordering. Productivity As we go in order, we calculate the least cost for that Productivity Non-word error Non-word error node: detection detection We add the cost of an arc to the cost of reaching the N-gram analysi N-gram analys node this arc originates from. Isolated-word erro Isolated-word erro correction We take the minimum of the costs calculated for all arcs correction Rule-based me Rule-based me pointing to a node and store it for that node. Similarity key technic Similarity key techn The key point is that we are storing partial results along the way, instead of recalculating everything, every time Grammar correction Grammar correctio we compute a new path. Computing with Syntax Computing with Sunta: Caveat empto Language and Computers Language and More on grammar correction Topic 4: Writer's aids Introduction Introduction Semantic errors = errors where the sentence structure Error causes sounds okay, but it doesn't really mean anything.

• e.g., They are leaving in about fifteen minuets to go to her house.

 \Rightarrow minuets and minutes are both plural nouns, but only one makes sense here

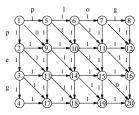
There are many different ways in which grammar correctors work, two of which we'll focus on:

- Bigram model (bigrams of words)
- Rule-based model

Computing edit distances

Adding costs to the arcs of the example graph

- We need to add the costs involved to the arcs.
- In the simplest case, the cost of deletion, insertion, and substitution is 1 each (and substitution with the same character is free).



 Instead of assuming the same cost for all operations, in reality one will use different costs, e.g., for the first character or based on the confusion probability.

Context-dependent word correction

Context-dependent word correction = correcting words based on the surrounding context.

- This will handle errors which are real words, just not the right one or not in the right form.
- Essentially a fancier name for a grammar checker = a mechanism which tells a user if their grammar is wrong.

Error causes Phonetic errors Difficult issues Tokenizatio Inflection Productivity Non-word error detection N-gram analysi Isolated-word erro correction Rule-based me Similarity key techniq Probal Minimum edit distance Grammar cor outing with Synta Caveat emptor

Caveat emptor

Computers

Topic 4:

Writer's aids

in a sentence: the order or form of words is incorrect, i.e., ungrammatical. Local syntactic errors: 1-2 words away ▶ e.g., The study was conducted mainly be John Black.

Syntactic errors = errors in how words are put together

Grammar correction—what does it correct?

- A verb is where a preposition should be. Long-distance syntactic errors: (roughly) 3 or more
- words away
 - e.g., The kids who are most upset by the little totem is aoina home early.
- Agreement error between subject kids and verb is

Language and

Topic 4 Writer's aid

Difficult issues Non-word error N-gram analysi

Computing with Supta Caveat emptor

Topic 4:

Writer's aids

Bigram grammar correctors	Language and Computers Topic 4: Writer's aids	Rule-based grammar correctors	Language and Computers Topic 4: Writer's aids	Beyond regular expressions	Language and Computers Topic 4: Writer's aids
 We can look at bigrams of words, i.e., two words appearing next to each other. Question: Given the previous word, what is the probability of the current word? e.g., given these, we have a 5% chance of seeing reports and a 0.001% chance of seeing report (these report cards). Thus, we will change report to reports But there's a major problem: we may hardly ever see these reports, so we won't know the probability of that bigram. (Partial) Solution: use bigrams of parts of speech. e.g., What is the probability of a noun given that the previous word was an adjective? 	Introduction Error causes Control ca	 We can write regular expressions to target specific error patterns. For example: To a certain extend, we have achieved our goal. Match the pattern some or certain followed by extend, which can be done using the regular expression some certain extend Change the occurrence of extend in the pattern to extent. Naber (2003) uses 56 such rules to build a grammar corrector which works nearly as well as that in commercial products. 	Introduction Error acuses Content area Content area Content area Content area Content	 But what about correcting the following: A baseball teams were successful. We should change A to Some, but a simple regular expression doesn't work because we don't know where the word teams might show up. A wildly overpaid, horrendous baseball teams were successful. (Five words later; change needed.) A player on both my teams was successful. (Five words later; no change needed.) We need to look at how the sentence is constructed in order to build a better rule. 	Introduction Encore autors Mathematicantes Mat
Syntax	Language and Computers Topic 4: Writer's aids	Linear order Linear order = the order of words in a sentence. 	Language and Computers Topic 4: Writer's aids	Constituency	Language and Computers Topic 4: Writer's aids
 Syntax = the study of the way that sentences are constructed from smaller units. There cannot be a "dictionary" for sentences since there is an infinite number of possible sentences: (7) The house is large. (8) John believes that the house is large. (9) Mary says that John believes that the house is large. There are two basic principles of sentence organization: Linear order Hierarchical structure (Constituency) 	Introduction Error causes Avade an ensystem Source errors Source errors Control intervent Source errors Source errors Source errors Control error Control error Control error Control error Control error Source errors Source errors Source errors Source errors Source errors Control errors Cont	 A sentence has different meanings based on its linear order. (10) John loves Mary. (11) Mary loves John. Languages vary as to what extent this is true, but linear order in general is used as a guiding principle for organizing words into meaningful sentences. Simple linear order as such is not sufficient to determine sentence organization though. For example, we can't simply say "The verb is the second word in the sentence." (12) I eat at really fancy restaurants. (13) Many executives eat at really fancy restaurants. 	Introduction Error causes Myseard minysings Posities errors Subsection Difficult issues Myseard errors Myseard errors Myseard	 What are the "meaningful units" of a sentence like Many executives eat at really fancy restaurants? Many executives really fancy really fancy restaurants at really fancy restaurants eat at really fancy restaurants We refer to these meaningful groupings as constituents of a sentence. 	Introduction Error causes Monards environ Monards environ Monards environ Monards environ Monards Monard
 Hierarchical structure Constituents can appear within other constituents. We can represent this in a bracket form or in a syntactic tree Constituents shown through brackets: [[Many executives] [eat [at [[really fancy] restaurants]]]] Constituents displayed as a tree: 	<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	 Categories We would also like some way to say that Many executives and really fancy restaurants are the same type of grouping, or constituent, whereas at really fancy restaurants seems to be something else. For this, we will talk about different categories Lexical Phrasal 	<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	 Lexical categories Lexical categories are simply word classes, or what you may have heard as parts of speech. The main ones are: verbs: eat, drink, sleep, nouns: gas, food, lodging, adjectives: quick, happy, brown, adverbs: quickly, happily, well, westward prepositions: on, in, at, to, into, of, determiners/articles: a, an, the, this, these, some, much, 	<section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>

				-	
Determining lexical categories	Language and Computers Topic 4:	Closed & Open classes	Language and Computers Topic 4:	Phrasal categories	Language an Computers Topic 4:
 How do we determine which category a word belongs to? Distribution: Where can these kinds of words appear in a sentence? e.g., Nouns like <i>mouse</i> can appear after articles ("determiners") like <i>some</i>, while a verb like <i>eat</i> cannot. Morphology: What kinds of word prefixes/suffixes can a word take? e.g., Verbs like <i>walk</i> can take a <i>ed</i> ending to mark them as past tense. A noun like <i>mouse</i> cannot. 	Writer's aids Introduction Encoduction May be an encoduction May	We can add words to some classes, but not to others. This also seems to correlate with whether a word is "meaningful" or just a function word = only meaning comes from its usage in a sentence. Open classes : new words can be easily added: • verbs • nouns • adjectives • adverbs Closed classes : new words cannot be easily added: • prepositions • determiners	Writer's aids Introduction Intr	 What about phrases? Can we assign them categories? We can also look at their distribution and see which ones behave in the same way. The joggers ran through the park. What other phrases can we put in place of <i>The joggers</i>? 	Writer's aids Introduction Error causes Cau
Phrasal categories (cont.)	Language and Computers Topic 4: Writer's aids	Building a tree	Language and Computers Topic 4: Writer's aids	Phrase Structure Rules	Language and Computers Topic 4: Writer's aids
 Susan students you most dogs some children a huge, lovable bear my friends from Brazil the people that we interviewed Since all of these contain nouns, we consider these to be noun phrases, abbreviated with NP. 	Introduction Error causes Kondead mitippings Nondead mitippings Nondead mitippings Nondead mitippings Nondead mitippings Nondead mitippings Non-word error Cottection Distances Nondead mitippings Nondead	Other phrases work similarly (S = sentence, VP = verb phrase, PP = prepositional phrase, AdjP = adjective phrase): NP VP Many executives eat PP at NP AdjP restaurants really fancy	Introduction Error causes Nonexer mithyings Phomeie error Scouledge problems Difficult issues Nonexer Nonexer Nonexer Decements Nonexer Decements Nonexer Decements Nonexer Nonexer Decements Nonexer Decements Nonexer Decements Nonexer Nonexer Nonexer Decements Nonexer No	 We can give rules for building these phrases. That is, we want a way to say that a determiner and a noun make up a noun phrase, but a verb and an adverb do not. Phrase structure rules are a way to build larger constituents from smaller ones. e.g., S → NP VP This says: A sentence (S) constituent is composed of a noun phrase (NP) constituent. (hierarchy) The NP must precede the VP. (linear order) 	Introduction Error causes Monotariangenetic Mono
Some other English rules $ ightarrow NP \rightarrow Det N$ (the cat, a house, this computer)	58/72 Language and Computers Topic 4: Writer's aids	Phrase Structure Rules and Trees	59/72 Language and Computers Topic 4: Writer's aids	Phrase Structure Rules in Practice	60 / Language ann Computers Topic 4: Writer's aids
 NP → Det AdjP N (<i>the happy cat, a really happy house</i>) For phrase structure rules, as shorthand parentheses are used to express that a category is optional. We thus can compactly express the two rules above as one rule: NP → Det (AdjP) N Note that this is different and has nothing to do with the use of parentheses in regular expressions. AdjP → (Adv) Adj (<i>really happy</i>) VP → V (<i>laugh, run, eat</i>) VP → V NP (<i>love John, hit the wall, eat cake</i>) VP → V NP NP (<i>give John the ball</i>) PP → P NP (<i>to the store, at John, in a New York minute</i>) NP → NP PP (<i>the cat on the stairs</i>) 	Introduction Fror causes From from causes From f	With every phrase structure rule, you can draw a tree for it.	Introduction Error causes Error	 Try analyzing these sentences and drawing trees for them, based on the phrase structure rules given above. The man in the kitchen drives a truck. That dang cat squeezed some fresh orange juice. The mouse in the corner by the stairs ate the cheese. 	Introduction Error causes Notice of manyors Productions Defined manyors Notice of the Notice of the

Properties of Phrase Structure Rules • generative = a schematic strategy that describes a set	Language and Computers Topic 4: Writer's aids	Context-free grammars	Language and Computers Topic 4: Writer's aids	Pushdown automata	Language and Computers Topic 4: Writer's aids
 of sentences completely. > potentially (structurally) ambiguous = have more than one analysis (14) We need more intelligent leaders. (15) Paraphrases: a. We need leaders who are more intelligent. b. Intelligent leaders? We need more of them! > hierarchical = categories have internal structure; they aren't just linearly ordered. > recursive = property allowing for a rule to be reapplied (within its hierarchical structure). e.g., NP → NP PP PP → P NP The property of recursion means that the set of potential sentences in a language is infinite. 	<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	 A context-free grammar (CFG) is essentially a collection of phrase structure rules. It specifies that each rule must have: a left-hand side (LHS): a single non-terminal element = (phrasal and lexical) categories a right-hand side (RHS): a mixture of non-terminal and terminal elements terminal elements A CFG tries to capture a natural language completely. Why "context-free"? Because these rules make no reference to any context surrounding them. i.e. you can't say "PP → P NP" when there is a verb phrase (VP) to the left. 	Introduction Error causes Madead minipairs Madead minipai	 Pushdown automaton = the computational implementation of a context-free grammar. It uses a stack (its memory device) and has two operations: push = put an element onto the top of a stack. pop = take the topmost element from the stack. This has the property of being Last In First Out (LIFO). So, when you have a rule like "PP → P NP", you push NP onto the stack and then push P onto it. If you find a preposition (e.g., <i>on</i>), you pop P off of the stack and now you know that the next thing you need is an NP. 	Introduction Error causes Manage interest Manage inter
Parsing	Language and Computers Topic 4: Writer's aids	Writing grammar correction rules	Language and Computers Topic 4: Writer's aids	Is this really how spell checkers work?	Language and Computers Topic 4: Writer's aids
 So, using these phrase structure (context-free) rules and using something like a pushdown automaton, we can get a computer to parse a sentence = assign a structure to a sentence. Do you parse top-down or bottom-up (or a mixture)? top-down: build a tree by starting at the top (i.e. S → NP VP) and working down the tree. bottom-up: build a tree by starting with the words at the bottom and working up to the top. There are many, many parsing techniques out there. 	Introduction Error causes Kooseade missions Koos	 So, with context-free grammars, we can now write some correction rules, which we will just sketch here. A baseball teams were successful. A followed by PLURAL NP: change A → The John at the taco. The structure of this sentence is NP PP, but that doesn't make up a whole sentence. We need a verb somewhere. 	Introduction Error causes Another and misyings Phonetic error Another and misyings Phonetic error Another and another Internation Internatio Internation Internation Internati	 As far as we know, yes, but: Many spell checkers are proprietary and the way they work is kept secret; we don't know how they work exactly, which hampers research and thereby progress. Others, such as aspell and ispell, are open source spell checkers, meaning that anyone can contribute to their further development, and see how they work, which makes it possible to understand exactly what they will and what they won't catch. (cf. http://aspell.sourceforge.net/ and http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html 	Introduction Error causes Phonetic error Noceedage problems Difficult issues Difficult issues Phonetic error Difficult issues Phonetic error Difficult issues Difficult issues Domonot error Difficult issues Domonot error Difficult issues Non-word error Di
 Dangers of spelling and grammar correction The more we depend on spelling correctors, the less we try to correct things on our own. But spell checkers are not 100% A study at the University of Pittsburgh found that students made more errors when using a spell checker! high SAT scores low SAT scores use checker 16 errors 17 errors no checker 5 errors 12.3 errors (cf., http://www.wired.com/news/business/0,1367,58058,00.html) 	67/72 Language and Computers Computers Tardoution Introduction Computers Computers <td>A Poem on the Dangers of Spell Checkers Michael Livingston Eye halve a spelling chequer It came with my pea sea. It plainly marques four my revue Miss steaks eye kin knot sea. Eye strike a key and type a word And weight four it two say Weather eye am wrong oar write It shows me strait a weigh. As soon as a mist ache is maid It nose bee fore two long And eye can put the error rite Its rare lea ever wrong. Eye have run this poem threw it I am shore your pleased two no Its letter perfect awl the weigh My chequer tolled me sew.</td> <td>Cashrage Cashra</td> <td> References The discussion is based on Markus Dickinson (2006). Writer's Aids. In Keith Brown (ed.): Encyclopedia of Language and Linguistics. Second Edition Elsevier. A major inspiration for that article and our discussion is Karen Kukich (1992): Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, pages 377–439. For a discussion of the confusion matrix, cf. Mark D. Kernighan, Kenneth W. Church and William A. Gale (1990). A spelling Correction Program Based on a Noisy Channel Model. In Proceedings of COLING-90. pp. 205–210. An open-source style/grammar checker is described in Daniel Naber (2003). A Rule-Based Style and Grammar Checker. Diploma Thesis, Universität Bielefeld. http://www.danielnaber.de/languagetool/ </td> <td>69/72 Ganguage and Computers Topic 4: Writer's aids Introduction Error causes Meabed manages Production Error causes Noneideproteins Difficult issues Todatases Difficult issues Todatases Non-word error detection Distants Name and stants Carpang with Synta C</td>	A Poem on the Dangers of Spell Checkers Michael Livingston Eye halve a spelling chequer It came with my pea sea. It plainly marques four my revue Miss steaks eye kin knot sea. Eye strike a key and type a word And weight four it two say Weather eye am wrong oar write It shows me strait a weigh. As soon as a mist ache is maid It nose bee fore two long And eye can put the error rite Its rare lea ever wrong. Eye have run this poem threw it I am shore your pleased two no Its letter perfect awl the weigh My chequer tolled me sew.	Cashrage Cashra	 References The discussion is based on Markus Dickinson (2006). Writer's Aids. In Keith Brown (ed.): Encyclopedia of Language and Linguistics. Second Edition Elsevier. A major inspiration for that article and our discussion is Karen Kukich (1992): Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, pages 377–439. For a discussion of the confusion matrix, cf. Mark D. Kernighan, Kenneth W. Church and William A. Gale (1990). A spelling Correction Program Based on a Noisy Channel Model. In Proceedings of COLING-90. pp. 205–210. An open-source style/grammar checker is described in Daniel Naber (2003). A Rule-Based Style and Grammar Checker. Diploma Thesis, Universität Bielefeld. http://www.danielnaber.de/languagetool/ 	69/72 Ganguage and Computers Topic 4: Writer's aids Introduction Error causes Meabed manages Production Error causes Noneideproteins Difficult issues Todatases Difficult issues Todatases Non-word error detection Distants Name and stants Carpang with Synta C