Starting to Process Social Media

Leon Derczynski
University of Sheffield
9 Dec 2013

„Infocommunication technologies and the society of future (FuturICT.hu)" TÁMOP-4.2.2.C-11/1/KONV-2012-0013

Functional utterances

Vowels

Velar closure: consonants

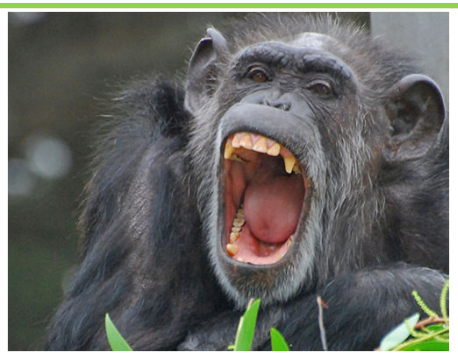
Speech

New modality: writing

Digital text

E-mail

Social media



Increased

machine-readable

information

„Infocommunication technologies and the society of future (FuturICT.hu)" TÁMOP-4.2.2.C-11/1/KONV-2012-0013

What resources do we have now?

Large, content-rich, linked, digital streams of human communication

We transfer knowledge via communication

Sampling communication gives a sample of human knowledge

''You've only done that which you can communicate''

The metadata (time – place – imagery) gives a richer resource:

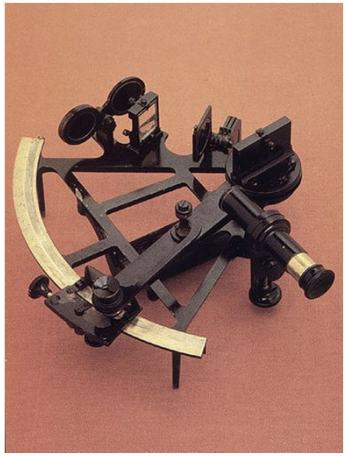
→ A sampling of human behaviour

What can we do with this resource?

Context increases the data's richness

Increased richness enables novel applications

Time and Place are interesting parts of message context



1.What kinds of applications are there?

2.What are the practical challenges?

Social media analysis

Ability to extract sequences of events*

Retrieve information on:

- Lifecycle of socially connected groups
- Analyse precursors to events, post-hoc

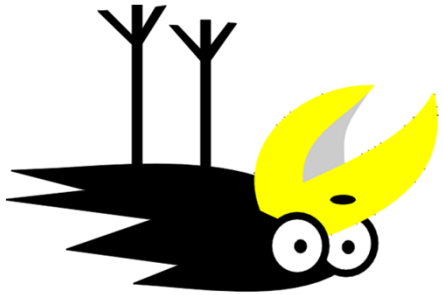
2008

2011

* Weikum et al. 2011: ''Longitudinal analytics on web archive data: It's about time'', *Proc. CIDR*

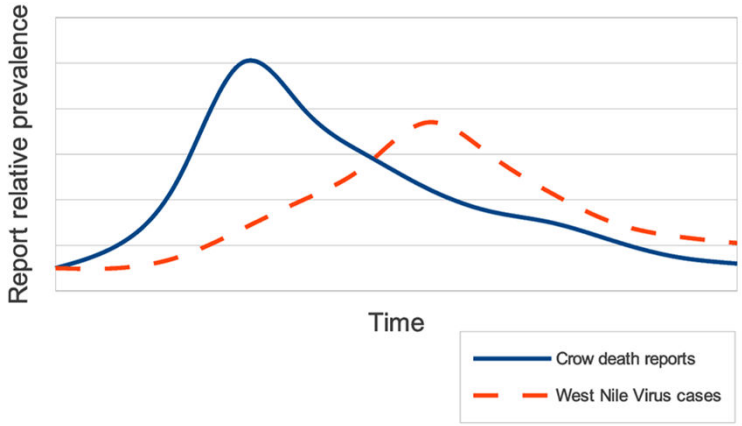
Social media analysis

Retrospective analyses into cause and effect



"There's a dead crow in my garden"

Social media mentions of dead crows predict WNV in humans *



* Sugumaran & Voss 2012: ''Real-time spatio-temporal analysis of West Nile Virus using Twitter Data'', *Proc. Int'l conferenc*

Features of Social Network Sites

- Open access; anyone can post
- No "editing" or approval
- Option to include extra data – URLs, photos
- Friendship relations between posters
  - Uni-directional: Twitter
  - Bi-directional: Facebook
- Profile metadata
- API for access
- Rapidly updated content

- **Summary: a rich, varied, fast-moving network of information**

Social Media = Big Data

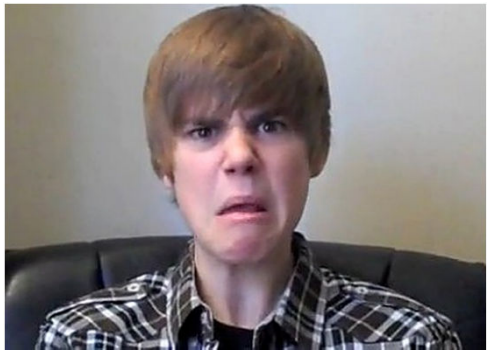
Gartner ''3V'' definition:

1.Volume
2.Velocity
3.Variety

High <u>volume</u> & <u>velocity</u> of messages:

Twitter has ~20 000 000 users per month
They write ~500 000 000 messages per day

Massive <u>variety</u>:
Stock markets;
Earthquakes;
Social arrangements;
… Bieber



Emerging search

Data emerging at high velocity:

185 000 documents per minute

Gives a high temporal density

Search over this info enables:

- Live coverage of events
- Realtime identification of emerging events *

*** Cohen at al. 2011: ''Computational journalism: A call to arms to database researchers'', *Proc. CIDR***

Challenges in analysing social media

We've seen:

- What social media is
- How social media data is represented

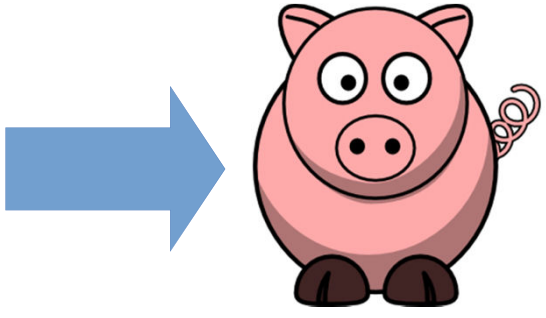
What are the challenges in doing NLP on social media data?

Why do traditional NLP models not work well?

Typical annotation pipeline

Text

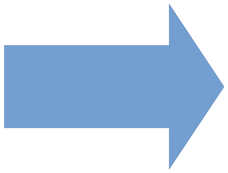
Language ID

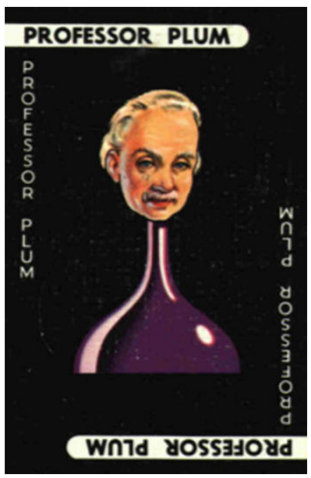


Tokenisation

PoS tagging

Typical annotation pipeline



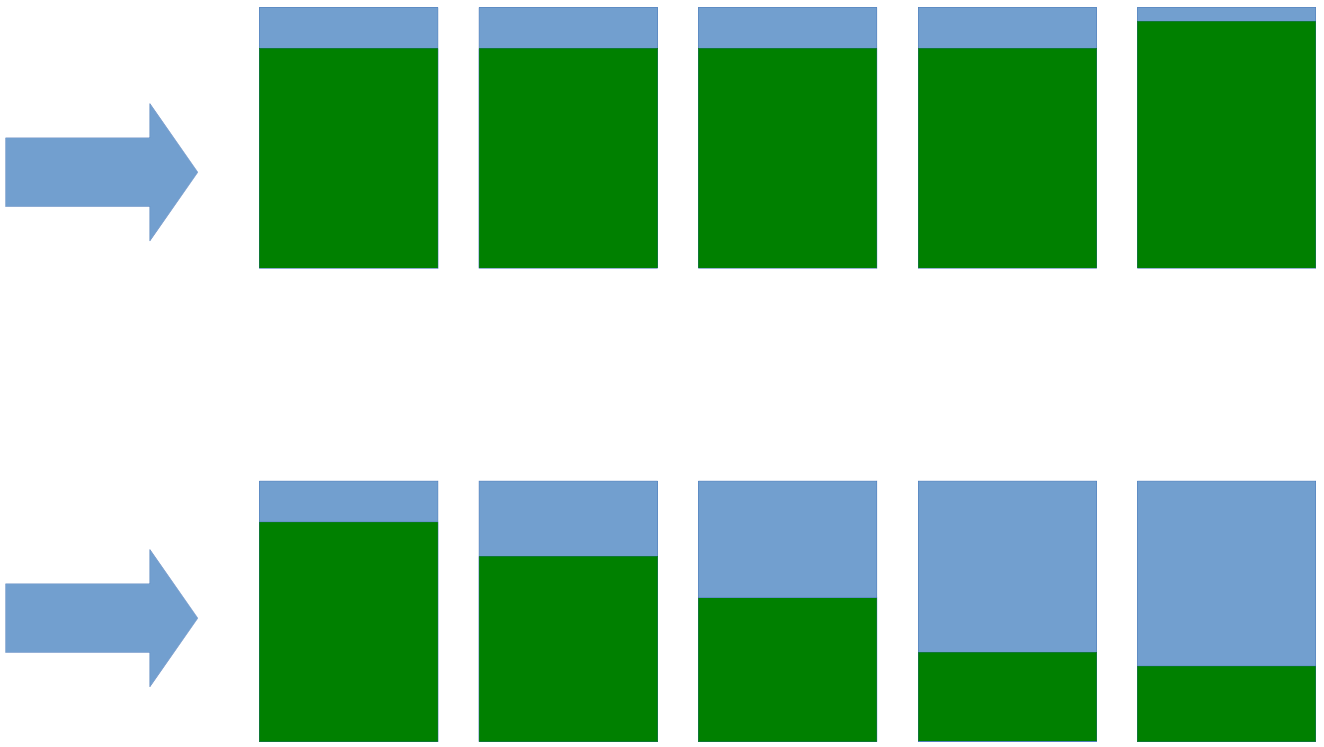


Named entity recognition

dbpedia.org/resource/.....
Michael_Jackson
Michael_Jackson_(writer)

Linking entities

Pipeline cumulative effect



Good performance is important at each stage
- not just the end

Language ID

The Jan. 21 show started with the unveiling of an impressive three-story castle fro

Newswire:

Microblog: LADY GAGA IS BETTER THE 5th TIME OH BABY(:

Language ID difficulties

General accuracy on microblogs: 89.5%

Problems include switching language mid-text:

je bent Jacques cousteau niet die een nieuwe soort heeft ontdekt, het is duidelijk, ze bedekk

New info in this format:

Metadata:

spatial information
linked URLs

Emoticons:

:) vs. ^_^
cu vs. 88

Accuracy when customised to genre: 97.4%

Tokenisation

General accuracy on microblogs: 80%

Goal is to convert byte stream to readily-digestible word chunks

Word bound discovery is a *critical* language acquisition task

The LIBYAN AID Team successfully shipped these broadcasting equipment to Misrata l

RT @JosetteSheeran: @WFP #Libya breakthru! We move urgently needed #food (wheat, flou

Tokenisation difficulties

Not curated, so typos

Improper grammar – e.g. apostrophe usage; live with it!

doesnt → doesnt
doesn't → does n't

Smileys and emoticons

I <3 you → I & lt ; you
This piece ;,,(so emotional → this piece ; , , (so emotional

Loss of information (sentiment)

Punctuation for emphasis

*HUGS YOU**KISSES YOU* → * HUGS YOU**KISSES YOU *

Words run together

Part of speech tagging

Goal is to assign words to classes (verb, noun etc)

General accuracy on newswire:

97.3% token, 56.8% sentence

General accuracy on microblogs:

73.6% token, **4.24**% sentence

Sentence-level accuracy important:

without whole sentence correct, difficult to extract syntax

Part of speech tagging difficulties

Many unknowns (i.e. tokens not seen in training data):

Music bands:
Soulja Boy | TheDeAndreWay.com in stores Nov 2, 2010
Places:
#LB #news: Silverado Park Pool Swim Lessons

Capitalisation way off
@thewantedmusic on my tv :) aka derek
last day of sorting pope visit to birmingham stuff out

Slang
~HAPPY B-DAY TAYLOR !!! LUVZ YA~

Orthographic errors
dont even have homwork today, suprising ?

Dialect
Shall we go out for dinner this evening?
Ey yo wen u gon let me tap dat

Named Entity Recognition

Goal is to find entities we might like to link

London Fashion Week grows up – but mustn't take itself too seriously. Onc

Newswire:

Microblog: Gotta dress up for london fashion week and party in style!!!

General accuracy on newswire: 89% F1

General accuracy on microblogs: 41% F1

Person mentions in news

Left context	Match	Right context
in dicated Atef, including	Douglas Feith	, the United States defence
, the group that killed	President Sadat	in 1981 as retribution for
. The current leader,	President Olusegun Obasanjo	, who recently came to
Kuwait, whose information minister	Sheikh Ahmed Fahed al-Sabah	met editors of local newspapers
The current defence minister,	Theophilus Danjuma	, has also been threatened
The three right-wing MPs,	Andrew Rosindell	(Romford), Andrew
Late on Wednesday night,	Justice Oputa	, who chairs the commission
the militarily-manoeuvred civilian elec...	President Obasanjo	in 1999 and is widely
after the mysterious death of	General Sani Abacha	in 1998.
have learnt that one of	Bin Laden	's closest and most senior
evidence confirms the involvement of	Osama bin Laden	in those attacks."
. He is one of	Bin Laden	's two most senior associates
for future civilian office.	General Buhari	took power in a 1983
$5m price on	Atef	's head and prosecutors have
Afghanistan. He was once	Bin Laden	's chief media adviser and
thinking in the Tory party	Iain Duncan Smith	has ordered three Tory MPs
club and the party,	David Maclean	, the Tory Chief Whip
Centre and the Pentagon.	Mohammed Atef	, who is thought to
are still very powerful.	General Babangida	supported the militarily-manoeuvred civ
sexual orientation or religion.	Mr Duncan Smith	's purge of the Monday
," he said.	Atef	, who is reported variously
of the late singer,	Fela Kuti	🔲 which took place while
field in Penn sylvania.	President Bush	included Atef in an order
. It is believed that	Mr Duncan Smith	intended to launch his crackdown

Person mentions in tweets

Left context	Match	Right context
i was your age ,	spencer	from iCarly was Crazy Steve
iCarly was Crazy Steve ,	Carly	was Megan and Josh was
bath , shut up ,	sam	's coming tomorrow and steve
. All are welcome ,	joe	included
. All are welcome ,	joe	included
teachers , chinese takeaways ,	gatt holly	, phil collins , the
takeaways , gatt holly ,	phil collins	, the skin of a
@GdnPolitics : RT AlJahom :	Blair	: " I'm gonna
Empls of the Month :	Deborah L	#Speech #Pathologist-Childrens
be the next Pope "	Brown	: " I won't
(via POPSUGAR)	Sarah Jessica Parker	and Gwen Stefani Wrap Up
and is smexy !!;)And	Chelsea Handler	is hilarious ! Finally got
him befnrjustthen about	kenny	signing his book but it
three kinds of reactions after	Ayodhya	verdict .
, Carly was Megan and	Josh	was fat . #damnteenquotes
sam 's coming tomorrow and	steve	and tanya will be round
coming tomorrow and steve and	tanya	will be round at 10am
photo caption contest- Nadal and	Novak	in the tub http://ow.ly/2G3Jh
) Sarah Jessica Parker and	Gwen Stefani	Wrap Up Another Successful New
#Pathologist-Childrens Rehab and	Patricia M	#Referral/#Auth #
Just casually stalking Cheryl AND	Dermot	tomorrow .... NO BIGGIE
did tweet him befnr	justthen	about kenny signing his book
Test : We just congratulated	Lindsay	an hour ago on h
the funny photo caption contest-	Nadal	and Novak in the tub

Genre Differences in Entity Types

Tweet-specific NER challenges

- Capitalisation is not indicative of named entities
- All uppercase, e.g. APPLE IS AWSOME
- All lowercase, e.g. all welcome, joe included
- All letters upper initial, e.g. 10 Quotes from Amy Poehler That Will Get You Through High School
- Unusual spelling, acronyms, and abbreviations
- Social media conventions:
- Hashtags, e.g. #ukuncut, #RusselBrand, #taxavoidance
- @Mentions, e.g. @edchi (PER), @mcg_graz (LOC), @BBC (ORG)

NER difficulties - summary

Rule-based systems get the bulk of entities (newswire 77% F1)

ML-based systems do well at the remainder (newswire 89% F1)

Small proportion of
difficult entities

Many complex issues

Using improved pipeline:
ML struggles, even with in-genre data: 49% F1
Rules cut through microblog noise: **80%** F1

NER on Facebook

Longer texts than tweets

Still has informal tone

MWEs are a problem!

- all capitalised:

Green Europe Imperiled as Debt Crises
Triggers Carbon Market Drop

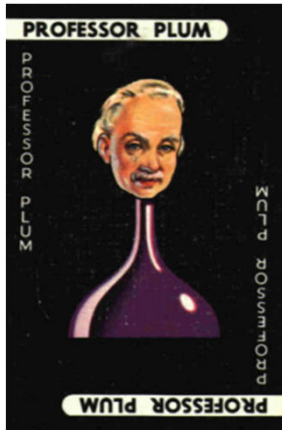
Difficult, though easier than Twitter

Maybe due to option of including more verbal context?

Entity linking

Goal is to find out which entity a mention refers to

''The murderer was Professor Plum, in
the Library, with the Candlestick!''



Which Professor Plum?

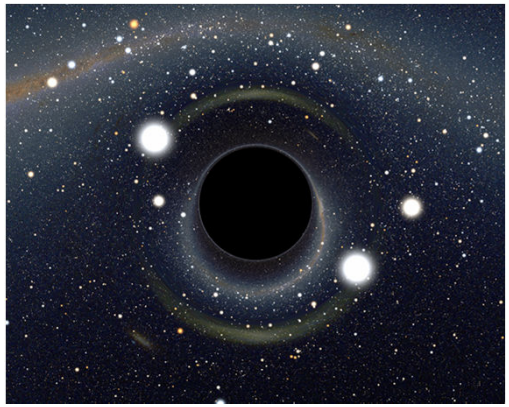
Disambiguation is through connecting text to the web of data
dbpedia.org/resource/Professor_Plum_(astrophysicist)

Word-level linking

Goal is to link an entity

Given:
The entity mention
Surrounding microblog context

No corpora exist for this exact task:
Two commercially produced
Company Policy says:
  ''no sharing''



How can we approach this key task?

Word-level linking performance

Dataset: Replab

Task is to determine relatedness-or-not

Six entities given

Few hundred tweets per entity

Detect mentions of entity in tweets

We disambiguate mentions to DBpedia / Wikipedia (easy to map)

General performance: F1 around 70

Word-level linking issues

NER errors

Missed entities damages / destroys linking

Specificity problems

Lufthansa Cargo
Lufthansa Cargo

Which organisation to choose?

Require good NER

Direct linking chunking reduces precision:

Apple trees in the home garden bit.ly/yOztKs

Pipeline NER does not mark Apple as entity here

Lack of disambiguation context is a problem!

Word-level linking issues

Automatic annotation:

Branching out from Lincoln park(LOC) after dark ... Hello "Russian Navy(ORG)", it's like the sa

Actual:

Branching out from Lincoln park after dark(ORG) ... Hello "Russian Navy (PROD)", it's like the

+

?

Whole pipeline: how to fix?

Common genre problems centre on mucky, uncurated text

Orth error

Slang

Brevity

Condensed

Non-Chicago punctuation..

Maybe clearing up this will improve performance?

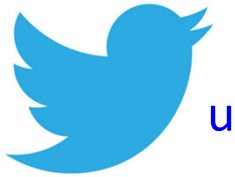
Normalisation

General solution for overcoming linguistic noise

How to repair?

1. Gazetteer (quick & dirty); or..
2. Noisy channel model

An honest, well-formed sentence



u wot m8 biber #lol

Task is to “reverse engineer'' the noise on this channel

What well-formed sentence would cause the observed tweet?

Brown clustering; double metaphone; auto orth correction

GATE for social media

GATE intro:

- Loading documents
- Annotations
- Annotation sets
- Annotation features
- Datastores
- Corpora
- Corpus pipelines
- Corpus quality
- Evaluation

Resources in GATE

- GATE treats most things as resources
  - Language resources (LR)
  - Processing resources (PR)
  - Visual resources (VR)

- Each kind of resource behaves in a different way
- These can be serialised and saved for sharing and reproducibility

Parameters

- Applications, LRs, and PRs all have various parameters which can be set either at load time (initialisation) or at run time.
- Parameters enable different settings to be used, e.g. case sensitivity
- **Initialisation Parameters** (set at load time**)** cannot be changed without reloading (these may be called “init parameters” for short)
- **Run time Parameters** can be changed between each application run
- Later you'll be able to experiment with setting parameters on resources and applications

Loading a document

- When GATE loads a document, it converts it into a special format for processing
- GATE can process documents in all kinds of formats: plain text, HTML, XML, PDF, Word etc.
- Documents have a markupAware parameter which is set to true by default: this ensures GATE will process any existing annotations such as HTML tags and present them as annotations rather than leaving them in the text.
- Documents can be exported in various formats or saved in a datastore for future processing within GATE

3. All about Annotations

- Introduction to annotations, annotation types and annotation sets
- Creating and viewing annotations

Annotations

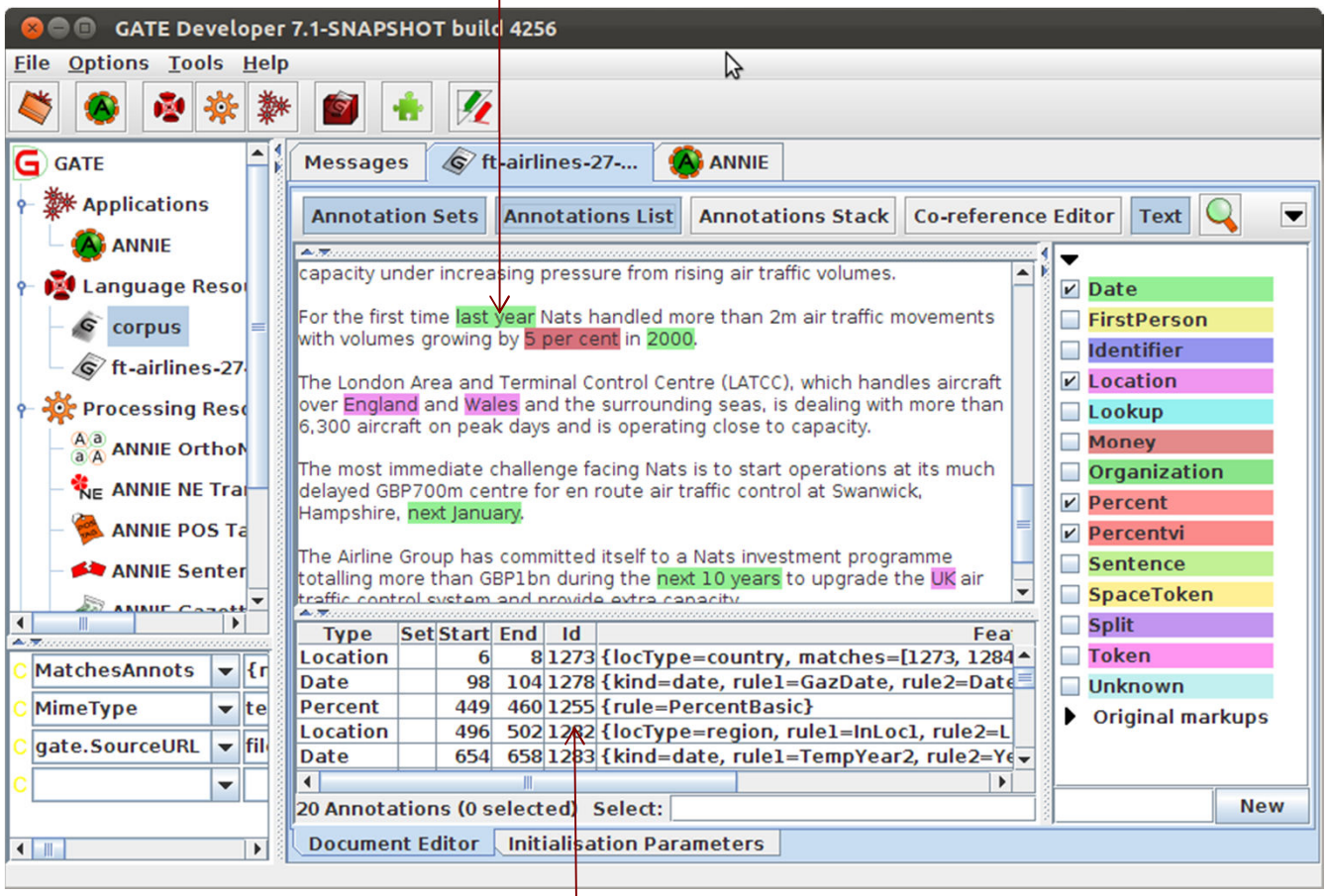
- The annotations associated with each document are a structure central to GATE.
- Each annotation consists of
- start and end offsets
- optionally a set of features associated with it
- each feature has a name and a value

Annotation Sets

- Annotations are grouped into sets, e.g. Default, Original Markups
- Each set can contain a number of annotation types, e.g. Person, Location etc.
- You can create and organise your annotation sets as you wish.
- It's useful to keep different sets for different tasks you may perform on a document, e.g. to separate the original HTML tags from your new annotations
- It's important to understand the distinction between annotation set, annotation type, and annotation
- This is best explained by looking at them in the GUI

Annotations

Date annotation



Annotations table

4. Documents and Corpora

- Creating and populating a corpus of documents in different ways

4. Documents and Corpora

- Documents and corpora are all language resources
- Documents can be created within GATE, or imported from other files and URLs
- GATE attempts to preserve any annotations in the existing document
  - XML markup
  - TEI
  - HTML
- Corpora consist of a set of documents
  - Can be created from already-imported docs
  - Can be automatically populated
- Try not to keep too much in memory at once..

5. Processing Resources and Plugins

- Loading processing resources and managing plugins

Processing Resources and Plugins

- Processing resources (PRs) are the tools that enable annotation of text. They implement algorithms. Typically this means creating or modifying annotations on the text.
- An application consists of any number of PRs, run sequentially over a corpus of documents
- A plugin is a collection of one or more PRs, bundled together. For example, all the PRs needed for IE in Arabic are found in the Lang_Arabic plugin.
  - A plugin may also contain language or visual resources, but you don't need to worry about that now!
- An application can contain PRs from one or more different plugins.
- In order to access new PRs, you need to load the relevant plugin

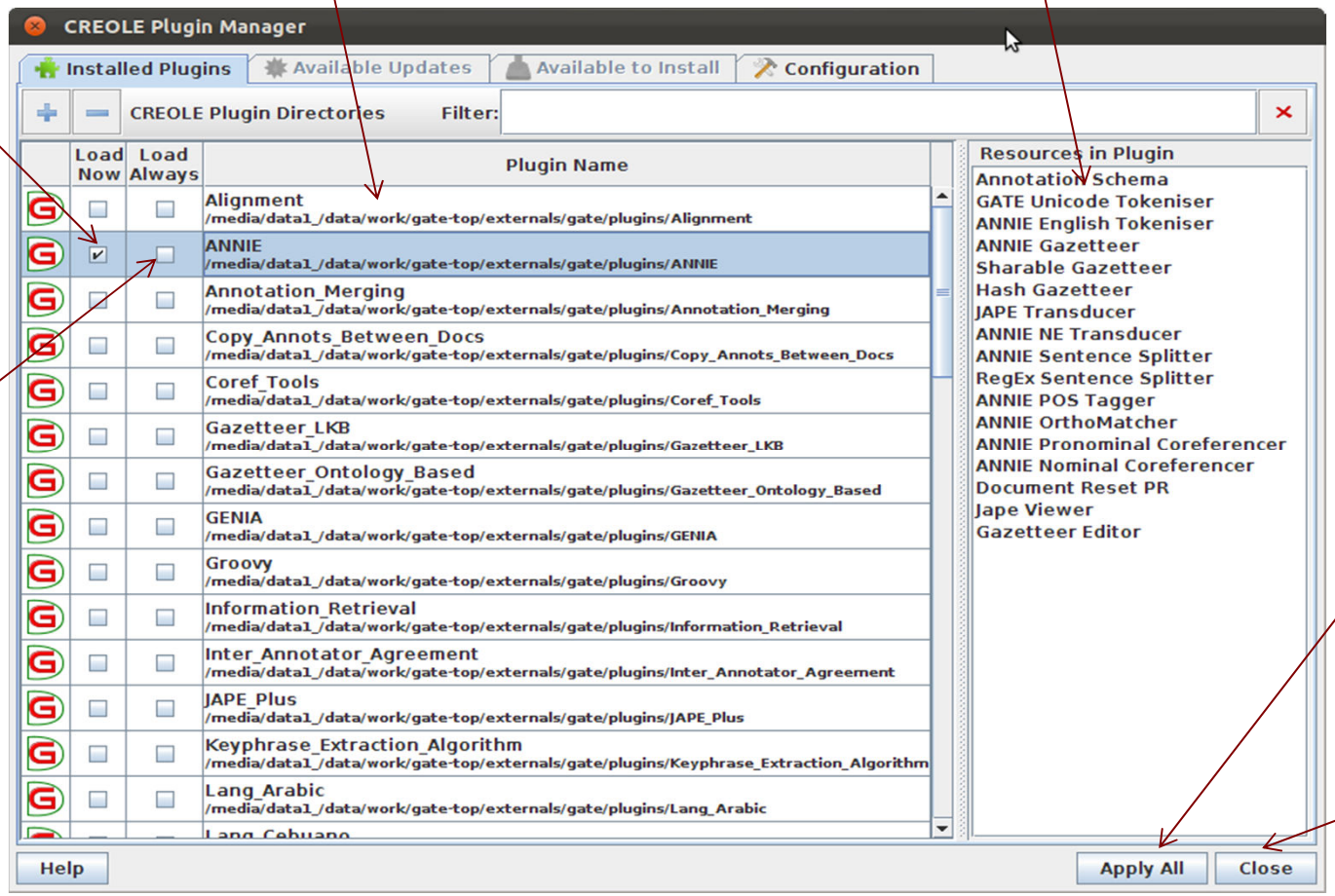
Plugins

Load the plugin for this session only

List of available plugins

Resources in the selected plugin

Load the plugin everytime GATE starts



Apply all the settings

Close plugins manager

6. Saving documents

- Using datastores
- Saving documents for use outside GATE

Types of datastores

There are 2 types of datastore:

- Serial datastores store data directly in a directory
- Lucene datastores provide a searchable repository with Lucene-based indexing

In this course, we consider the first type – but when you have advanced search requirements, a Lucene datastore can help

Saving documents outside GATE

- Datastores can only be used inside GATE, because they use some special GATE-specific format
- If you want to use your documents outside GATE, you can save them in 2 ways:
  - as standoff markup, in a special GATE representation
  - as inline annotations (preserving the original format)
- Both formats are XML-based. However “save as xml” refers to the first option, while “save preserving format” refers to the second option.

Information Extraction with GATE

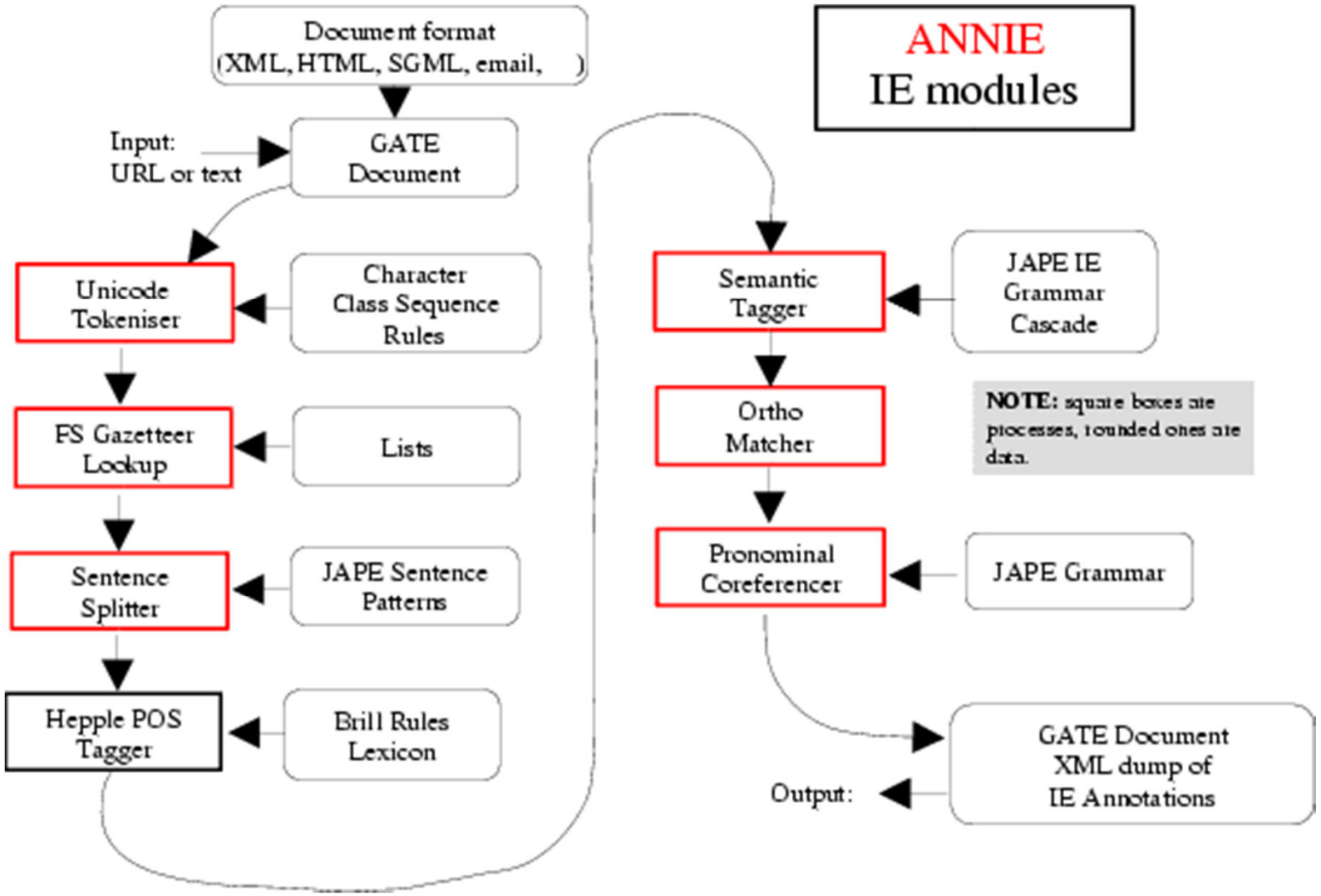
Nearly New Information Extraction

- ANNIE is a ready made collection of PRs that performs IE on unstructured text.
- For those who grew up in the UK, you can think of it as a Blue Peter-style “here's one we made earlier”.
- ANNIE is “nearly new” because
- It was based on an existing IE system, LaSIE
- We rebuilt LaSIE because we decided that people are better than dogs at IE
- Being 10 years old, it's not really new any more

What's in ANNIE?

- The ANNIE application contains a set of core PRs:
- Tokeniser
- Sentence Splitter
- POS tagger
- Gazetteers
- Named entity tagger (JAPE transducer)
- Orthomatcher (orthographic coreference)
- There are also other PRs available in the ANNIE plugin, which are not used in the default application, but can be added if necessary
- NP and VP chunker

Core ANNIE components



Let's look at the PRs

- Each PR in the ANNIE pipeline creates some new annotations, or modifies existing ones
- Document Reset → removes annotations
- Tokeniser → Token annotations
- Gazetteer → Lookup annotations
- Sentence Splitter → Sentence, Split annotations
- POS tagger → adds category features to Token annotations
- NE transducer → Date, Person, Location, Organisation, Money, Percent annotations
- Orthomatcher → adds match features to NE annotations

Tokeniser

- Tokenisation based on Unicode classes
- Declarative token specification language
- Produces Token and SpaceToken annotations with features orthography and kind
- Length and string features are also produced
- Rule for a lowercase word with initial uppercase letter

```
"UPPERCASE_LETTER" LOWERCASE_LETTER"* >
  Token; orthography=upperInitial; kind=word
```

ANNIE English Tokeniser

- The English Tokeniser is a slightly enhanced version of the Unicode tokeniser
- It comprises an additional JAPE transducer which adapts the generic tokeniser output for the POS tagger requirements
- It converts constructs involving apostrophes into more sensible combinations
- don’t → do + n't
- you've → you + 've

POS tagger

- ANNIE POS tagger is a Java implementation of Brill's transformation based tagger
- Previously known as **Hepple Tagger** (you may find references to this and to **heptag**)
- Trained on WSJ, uses Penn Treebank tagset
- Default ruleset and lexicon can be modified manually (with a little deciphering)
- Adds category feature to Token annotations
- Requires Tokeniser and Sentence Splitter to be run first

Morphological analyser

- Not an integral part of ANNIE, but can be found in the Tools plugin as an “added extra”
- Flex based rules: can be modified by the user (instructions in the User Guide)
- Generates “root” feature on Token annotations
- Requires Tokeniser to be run first
- Requires POS tagger to be run first if the considerPOSTag parameter is set to true

Gazetteers

- Gazetteers are plain text files containing lists of names (e.g rivers, cities, people, …)
- The lists are compiled into Finite State Machines
- Each gazetteer has an index file listing all the lists, plus features of each list (majorType, minorType and language)
- Lists can be modified either internally using the Gazetteer Editor, or externally in your favourite editor (note that the new Gazett`eer editor replaces the old GAZE editor you may have seen previously)
- Gazetteers generate Lookup annotations with relevant features corresponding to the list matched
- Lookup annotations are used primarily by the NE transducer
- Various different kinds of gazetteer are available: first we'll look at the default ANNIE gazetteer

Editing gazetteers outside GATE

- You can also edit both the definition file and the lists outside GATE, in your favourite text editor
- If you choose this option, you will need to reinitialise the gazetteer in GATE before running it again
- To reinitialise any PR, right click on its name in the Resources pane and select “Reinitialise”

List attributes

- When something in the text matches a gazetteer entry, a Lookup annotation is created, with various features and values
- The ANNIE gazetteer has the following default feature types: majorType, minorType, language
- These features are used as a kind of classification of the lists: in the definition file features are separated by ":"
- For example, the "city" list has a majorType "location" and minorType "city", while the "country" list has "location" and "country" as its types
- Later, in the JAPE grammars, we can refer to all Lookups of type location, or we can be more specific and refer just to those of type "city" or type "country"

Ontologies in IE

- A typical way to use an ontology in IE is to create a gazetteer from names and labels in the ontology, and use this to annotate entities with IDs (URIs) from the ontology
- GATE includes several tools to help with this, including a basic ontology viewer and editor, several ontology backed gazetteers, and the ability to refer to ontology classes in grammars
- The extra exercises includes an example for you to try, a simple demo application that creates a gazetteer from a SPARQL endpoint, adds entity annotations, and then adds further information to the entities, from the ontology

NE transducer

- Gazetteers can be used to find terms that suggest entities
- However, the entries can often be ambiguous
  - "May Jones" vs "May 2010" vs "May I be excused?"
  - "Mr Parkinson" vs "Parkinson's Disease"
  - "General Motors" vs. "General Smith"
- Handcrafted grammars are used to define patterns over the Lookups and other annotations
- These patterns can help disambiguate, and they can combine different annotations, e.g. Dates can be comprised of day + number + month
- NE transducer consists of a number of grammars written in the JAPE language

Using co-reference

- Different expressions may refer to the same entity
- Orthographic co-reference module (orthomatcher) matches proper names and their variants in a document
- [Mr Smith] and [John Smith] will be matched as the same person
- [International Business Machines Ltd.] will match [IBM]

Orthomatcher PR

- Performs co-reference resolution based on orthographical information of entities
- Produces a list of annotation IDs that form a co-reference “chain”
- List of such lists stored as a document feature named “MatchesAnnots”
- Improves results by assigning entity type to previously unclassified names, based on relations with classified entities
- May not reclassify already classified entities
- Classification of unknown entities very useful for surnames which match a full name, or abbreviations,
e.g. “Bonfield” <Unknown> will match “Sir Peter Bonfield” <Person>
- A pronominal PR is also available

Language plugins

- Language plugins contain language-specific PRs, with varying degrees of sophistication and functions for:
- Arabic
- Cebuano
- Chinese
- Hindi
- Romanian
- There are also various applications and PRs available for French, German and Italian
- These do not have their own plugins as they do not provide new kinds of PR
- Applications and individual PRs for these are found in gate/plugins directory: load them as any other PR
- More details of language plugins in user guide

Building a language-specific application

- The following PRs are largely language-independent:
- Unicode tokeniser
- Sentence splitter
- Gazetteer PR (but do localise the lists!)
- Orthomatcher (depending on the nature of the language)

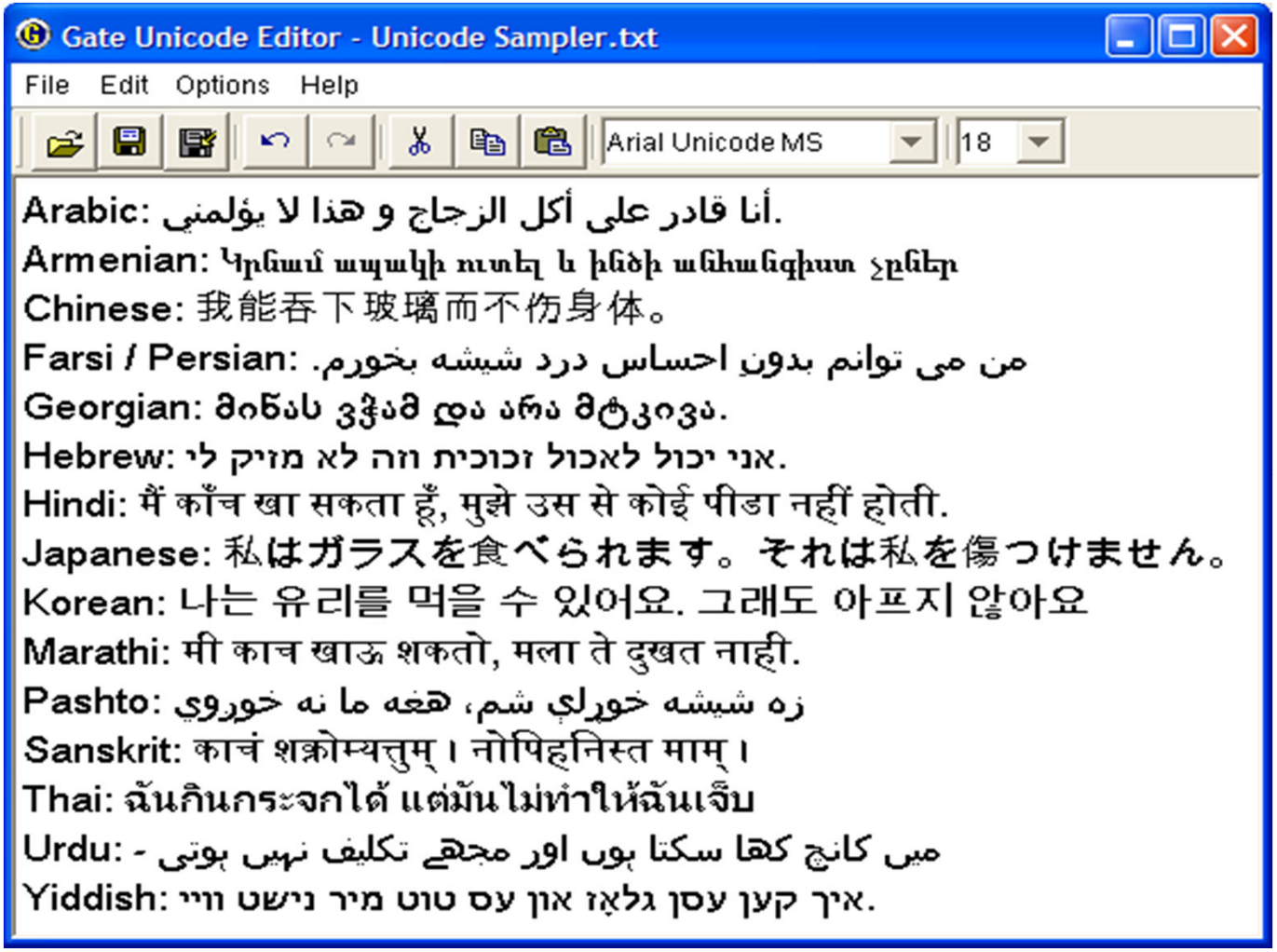
- Other PRs will need to be adapted (e.g. JAPE transducer) or replaced with a language-specific version (e.g. POS tagger)
- It can be helpful to consider “social media text” as its own separate language, for some pipeline stages

Useful Multilingual PRs

- Stemmer plugin
- Consists of a set of stemmer PRs for: Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, Swedish
- Requires Tokeniser first (Unicode one is best)
- Language is init-time param, which is one of the above in lower case
- Stanford tools
- Tokeniser, PoS tagger, NER
- TreeTagger
- a language-independent POS tagger which supports English, French, German and Spanish in GATE

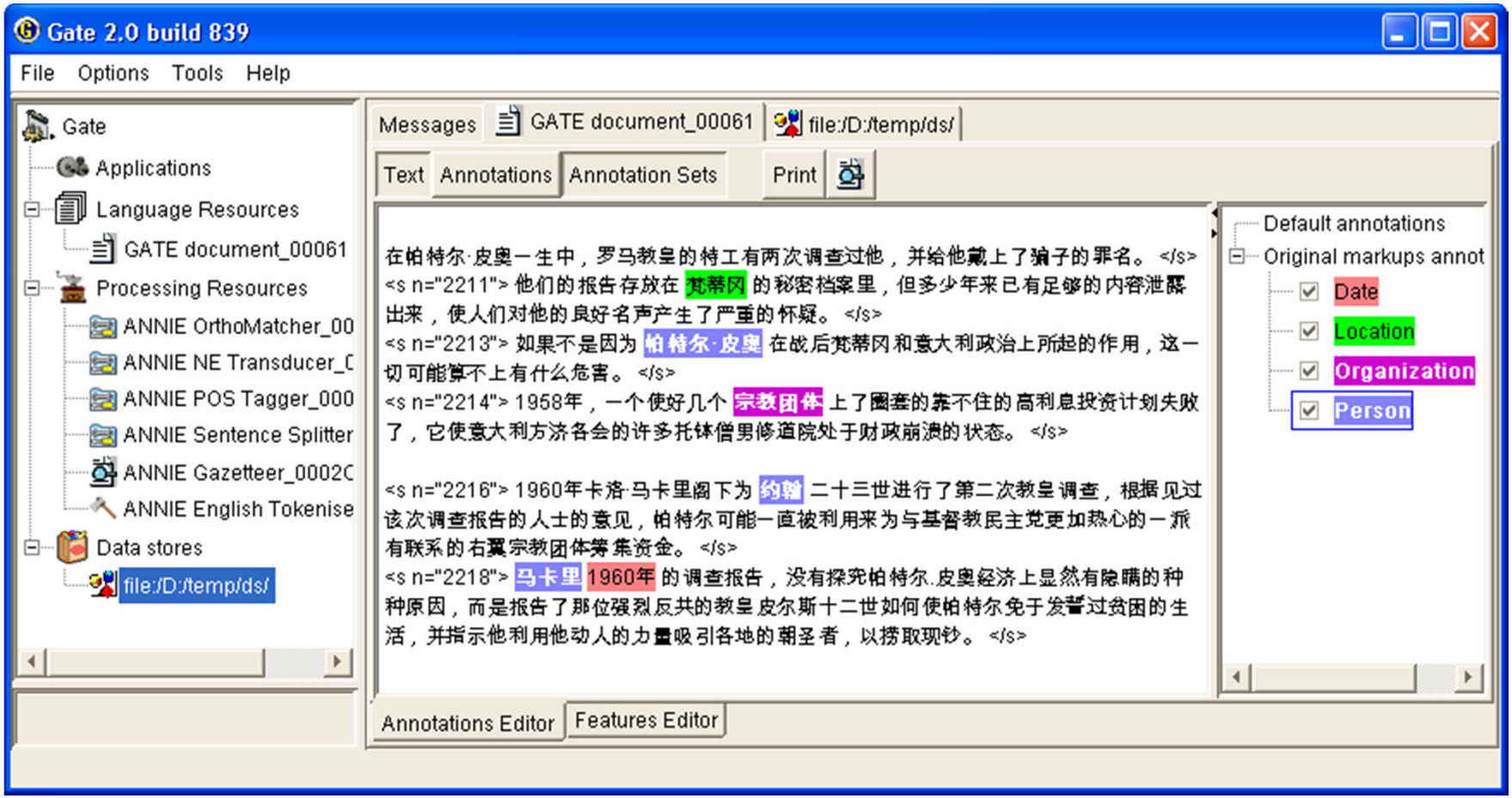
Displaying multilingual data

GATE uses standard (and imperfect) Java rendering engine for displaying text in multiple languages.



Displaying multilingual data

All visualisation and editing tools use the same facilities



Annotating social media

- Very few resources are available for social media text
  - No syntactic or dependency parsed data, no frames, no semantic roles
  - What corpora we do have are small (< 10k docs)
- There is no agreed standard for many parts of annotation:
  - Tokenisation: PTB, CMU
  - PoS tagset: PTB, CMU, Universal
  - Entity types: ACE, ACE+product, Ritter
- Reasonable chance that to do something new with social media NLP, you will need to annotate

Before you start annotating...

- You need to think about annotation guidelines
- You need to consider what you want to annotate and then to define it appropriately
- With multiple annotators it's essential to have a clear set of guidelines for them to follow
- Consistency of annotation is really important for a proper evaluation

Annotation Guidelines

- People need clear definition of what to annotate in the documents, with examples
- Typically written as a guidelines document
- Piloted first with few annotators, improved, then “real” annotation starts, when all annotators are trained
- Annotation tools may require the definition of a formal DTD (e.g. XML schema)
- What annotation types are allowed
- What are their attributes/features and their values
- Optional vs obligatory; default values

Evaluation

“We didn’t underperform. You overexpected.”

Performance Evaluation

2 main requirements:

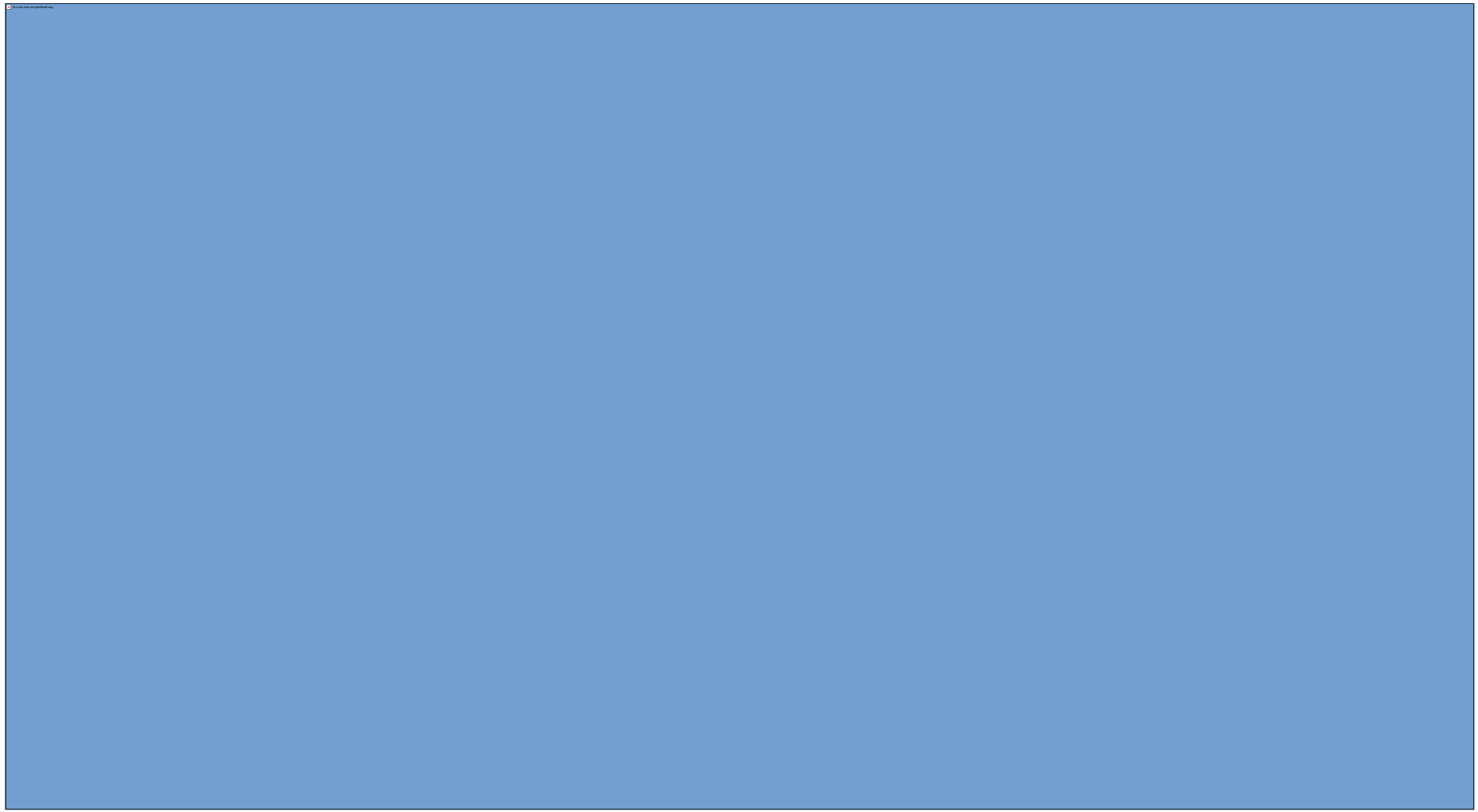
- **Evaluation metric**: mathematically defines how to measure the system's performance against human-annotated gold standard

- **Scoring program**: implements the metric and provides performance measures
  - For each document and over the entire corpus
  - For each type of annotation

A Word about Terminology

- Different communities use different terms when talking about evaluation, because the tasks are a bit different.
- The IE community usually talks about "correct", "spurious" and "missing"
- The IR community usually talks about "true positives", "false positives" and "negatives". They also talk about "false negatives", but you can ignore those.
- Some terminologies assume that one set of annotations is correct ("gold standard")
- Other terminologies do not assume one annotation set is correct
- When measuring inter-annotator agreement, there is no reason to assume one annotator is more correct than the other

Terminology Comparison



Measuring success

- In IE, we classify the annotations produced in one of 4 ways:
- **Correct** = things annotated correctly

  e.g. annotating “Hamish Cunningham” as a Person
- **Missing** = things not annotated that should have been

  e.g. not annotating “Sheffield” as a Location
- **Spurious** = things annotated wrongly

  e.g. annotating “Hamish Cunningham” as a Location
- **Partially correct** = the annotation type is correct, but the span is wrong

  e,g, annotating just “Cunningham” as a Person (too short) or annotating “Unfortunately Hamish Cunningham” as a Person (too long)

Summary:

Social media is rich
Social media is powerful
Social media is hard to process

This afternoon:

Practical GATE introduction

Thank you for your attention!

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013).