

# Intelligent Text Processing at Macquarie's Centre for Language Technology

Robert Dale

Robert.Dale@mq.edu.au

# Projects

---

## Established:

- **GainSpring:** automatic content extraction from news about companies [CMCRC]
- **DANTE:** temporal expression recognition and normalisation [DSTO]

## New:

- **Cross-document co-reference** [DSTO, UN]
- **Sentiment analysis in headlines** [Macq–UNSW]

[See other presenters for other relevant Macquarie projects]

# GainSpring

---

- **Funded by Capital Markets Co-operative Research Centre**
- **Adding value by finding information in unstructured data and repackaging this for easier and faster consumption**
- **Initial domain: ASX Company Announcements, 100000+ documents per year delivered via a web browser**
- **Extended to Hong Kong and London Stock Exchanges**

# GainSpring Components

---

- Text categorisation
- Named entity recognition
- Event recognition
- 'Best line' summarisation
- Hyperlinking of document space

[Home](#) > [Demo](#) > Recent Announcements

ASX Code	Summary	Published	Sensitive
CBA	Change in substantial holding for PWR	2000-12-29 16:52:35	
PWR	Change in substantial holding from CBA	2000-12-29 16:52:35	
PWR	Change in substantial holding from CBA	2000-12-29 16:52:35	
CBA	Change in substantial holding for HPX	2000-12-29 16:46:28	
HPX	<a href="#">Change in substantial</a>	Commonwealth 16:46:28	

[more announcements...](#)**Shareholding Company:**Bank of  
Australia  
(CBA)**Issuing Company:**Hpal Limited  
(HPX)**Date Of Transaction:**

2000-12-28

**Change:**

↑ from  
6,300,000  
(5.67) to  
7,778,500  
(7.00) ordinary  
shares

GainSpring is a project of the Capital Markets CRC

+61 2 9233 7999 | Level 2, 9 Castlereagh St, Sydney NSW 2006


[Home](#) > [Demo](#) > [Company details](#)

## Latest Director's transactions

Directors	Type	Shares	Date Notified	Previous Notification	Announcement
<a href="#">David Victor Murray</a>	Change	↓ 44,372	2000-07-01	1998-04-18	<a href="#">Announcement</a>
<a href="#">John Theodore Ralph</a>	Change	↓ 2,605	2000-04-14	1999-09-30	<a href="#">Announcement</a>
<a href="#">John Theodore Ralph</a>	Change	↓ 2,605	2000-04-14	1999-09-30	<a href="#">Announcement</a>
<a href="#">John Michael Schubert</a>	Change	↓ 6,476	2000-04-14	1999-11-19	<a href="#">Announcement</a>
<a href="#">Norman Ross Adler</a>	Change	↓ 6,222	2000-04-14	1999-09-30	<a href="#">Announcement</a>
<a href="#">Barbara Kay Ward</a>	Change	↓ 1,837	2000-04-14	1999-09-30	<a href="#">Announcement</a>
<a href="#">Frank Joseph Swan</a>	Change	↓ 1,922	2000-12-04	1999-09-30	<a href="#">Announcement</a>
<a href="#">Fergus Denis Ryan</a>	Change	↓ 4,000	2000-04-14		<a href="#">Announcement</a>
<a href="#">Anna Christina Booth</a>	Change	↓ 1,131	2000-04-13	1999-09-30	<a href="#">Announcement</a>
<a href="#">Warwick Gordon Kent</a>	Other	↓ 6,237			<a href="#">Announcement</a>
<a href="#">Warwick Gordon Kent</a>	Other	↓ 6,237			<a href="#">Announcement</a>

# Projects

---

## Established:

- GainSpring: automatic content extraction from news about companies [CMCRC]
- **DANTE: temporal expression recognition and normalisation [DSTO]**

## New:

- Cross-document co-reference [DSTO, UN]
- Sentiment analysis in headlines [Macq–UNSW]

[See other presenters for other relevant Macquarie projects]

# DANTE

---

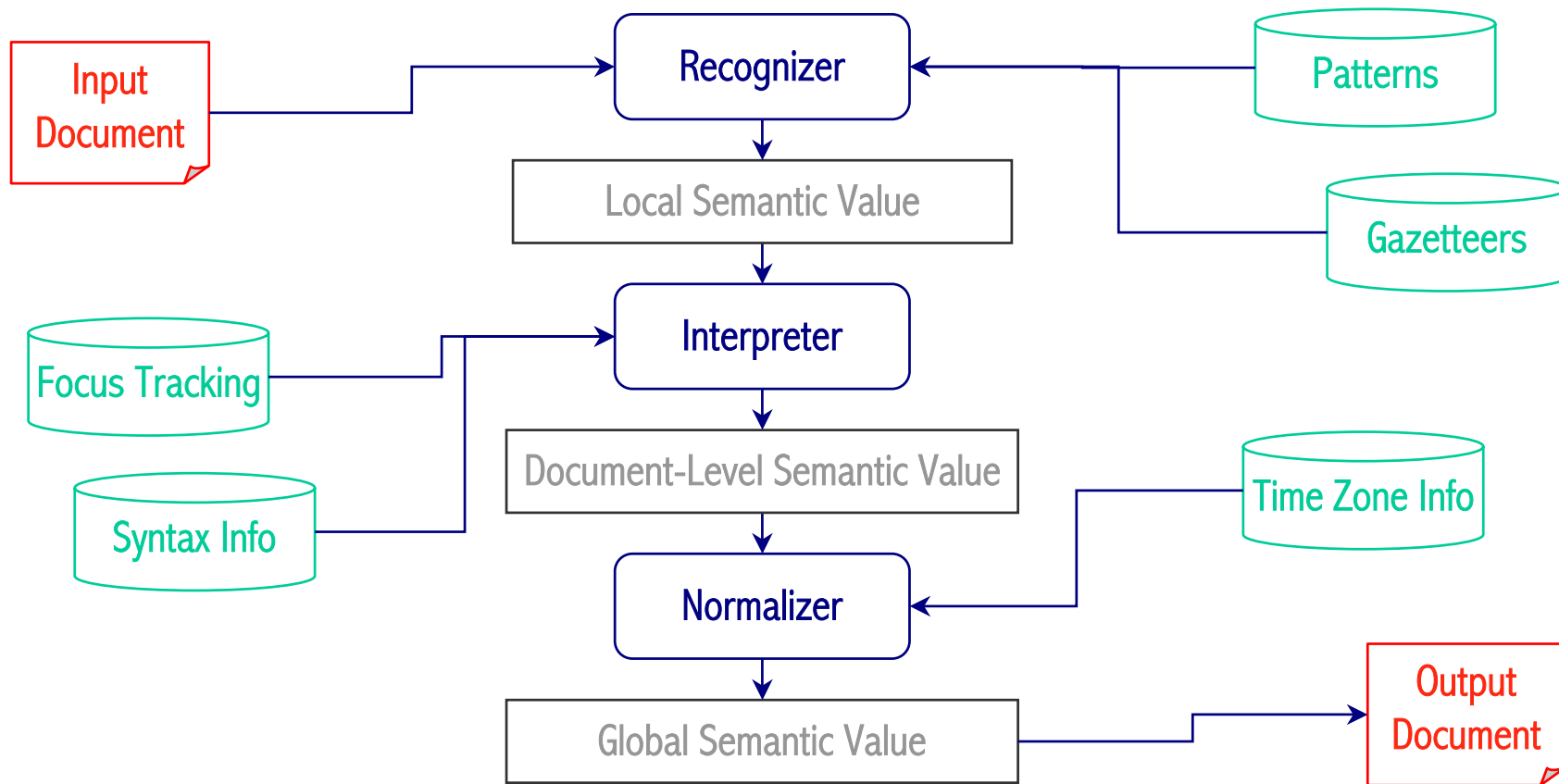
- **Funded by DSTO**
- **Time expression normalisation: given a temporal expression in a document, annotate this with an ISO date–time object, normalised to GMT time.**

# DANTE Examples

---

- The team arrived at 2006-05-04 1000 hours CST.  
...<TIMEX VAL="2006-05-04T00:30GMT">...
- The team arrived yesterday.  
...<TIMEX VAL="2006-05-03">...
- The team arrived at 5pm yesterday.  
...<TIMEX VAL="2006-05-03T17:00">...
- The team arrived last November.  
...<TIMEX VAL="2005-11">...

# DANTE's Architecture



# Projects

---

## Established:

- GainSpring: automatic content extraction from news about companies [CMCRC]
- DANTE: temporal expression recognition and normalisation [DSTO]

## New:

- **Cross-document co-reference [DSTO, UN]**
- Sentiment analysis in headlines [Macq–UNSW]

[See other presenters for other relevant Macquarie projects]

# Cross-Document Coreference

---

- **Funded by the DSTO**
- **Tracking named entities across multiple new sources**
- **Goal:**
  - **Using circumstantial evidence for identity: working out that two aliases refer to the same named entity on the basis of name-external evidence**
- **Preliminary results show name-external evidence based on co-occurring named entities provides best results**

# Cross-Document Coreference

---

- **New PhD project: Alex Rafalovitch**
- **Context:**
  - **Official Documents System of the United Nations (<http://documents.un.org/>): documents in all six official languages**
- **Goal:**
  - **Automatically extract and compile requests made to specific entities by various organs across a large number of distinct agendas**

# Projects

---

## Established:

- GainSpring: automatic content extraction from news about companies [CMCRC]
- DANTE: temporal expression recognition and normalisation [DSTO]

## New:

- Cross-document co-reference [DSTO, UN]
- **Sentiment analysis in headlines [Macq–UNSW]**

[See other presenters for other relevant Macquarie projects]

# Sentiment Analysis

---

- **Funded by Macq and UNSW internal grants schemes**
- **Goal:**
  - **See if we can learn to categorise sentiment expressed towards companies in newspaper stories**
- **Status:**
  - **Initial pilot annotation in progress**
- **Target:**
  - **Annotation of approximately 7500 stories (headlines and first paras) by three annotators**

# Sentiment Analysis: Example Text

---

**Telstra float may not end the pain**

**David Crowe**

**26 August 2006**

**Australian Financial Review**

**The government thinks it has an agreement to stop Telstra from railing about regulation. But the bad blood is not likely to end.**

**Is John Howard's Telstra nightmare finally over? Don't bet on it. The decision to proceed with an \$8 billion Telstra share offer settles the immediate question over the government's 51.8 per cent stake in the company. But the Prime Minister has made a fundamental decision about his policy on Telstra, and it is certain to pit the company and the government against each other for the long term. ...**

# Sentiment Analysis Categories

---

Positive	Straits flush after record profit result
Positive	Tattersalls upgrade restores confidence
Slightly Positive	Healthscope digests takeovers
Slightly Positive	Straits boosts production
Neutral	Veteran to retire from BHP Billiton
Neutral	BHP digs deep for diamonds
Slightly Negative	War of words delays Telstra ads
Slightly Negative	Idol 3 may be a lemon for Ten
Negative	Earnings downgrade to tell on Telstra bonds
Negative	Telstra falling behind the game

# Personnel

---

- **GainSpring: Li Lei, Hugo De Vries, Tim Yeates, Michele Wong**
- **DANTE: Pawel Mazur**
- **Cross-document co-reference: Pawel Mazur, Alex Rafalovitch**
- **Sentiment Analysis: Andrew Ferguson**

# Why is this Relevant to Next Generation Search?

---

- We need to move beyond bags of words
- But:
  - Full-blown analysis of NL text probably unachievable
  - I'm sceptical about human annotation for the semantic web
- So:
  - Use intelligent text processing to meta-tag documents with named entities, key events, time coding, and sentiment