Learner Corpus Research

Bergen, 27-29 September 2013

Comparing French/Spanish L1 transfers in two English learner corpora: the case of indexicals *it*, *this* and *that*

Thomas Gaillat – Université Paris-Diderot – Université de Rennes 1
CLILLAC-ARP EA 3967

Research question

How do (French/Spanish) learners of English use pro-forms (*it*, *this*, *that*)?

Outline

1. Introduction
2. Theoretical background
3. Method
4. Results and analysis
5. Conclusion and outlook

Introduction

- Positive transfers in SLA (Ellis 1996) for pronouns
- Identifying the forms: functional realisation taken into account
- Functional approach with Native and Non-native corpora :

  ► Measure differences in use between NNS & NS

Functional background

- *This* and *that* part of referential system. Fluctuating referential function: deictics
- Common functions : pro-forms, determiners, adverbials - homonymic *this* and *that*
- Micro-systems involving *this* & *that*
    - pro-forms compete with pronoun *it*
    - Determiners compete with *the*
- Functional analysis – function form approach (Ellis 2005)

Semantic background

- Different functions but same deictic-anaphoric value
    - Referring to a discourse entity
    - Distinction *deictic vs. anaphoric*: New or already existing (Cornish 1999)
    - Exophoric & endophoric reference (Halliday & Hassan 1976)
    - Speaker's sphere (Frazer & Joly, 1979)
        - *This* – speaker's sphere
        - *That* – outside the speaker's sphere
- *It* to simply identify an entity as known (Biber et al. 1999)

Tagset background

- Penn Treebank tagset
  - No distinctions between pro-forms and other uses of *it:* Empty subj/obj; Anticipatory subj/obj; Subject in clefts (Biber *et al.* 1999)
  - No distinction for pro-forms and determiner uses of *this* and *that* – one tag: DT
- Need for introduction of distinction

Method (1/3)

- Native corpus: ICE-GB – several categories of texts
  - 3 subsets: oral (general), written (general), written (student essays)
- Learner corpora
  - NOCE (347 871 words and signs) – Spanish students –written – argumentative essays (Diaz Negrillo 2004)
  - Diderot-LONGDALE (94 536 words and signs) –24 French students. Longitudinal: 3 years- Free speech oral expression about personal experience (Meunier et al. 2008)

Method (2/3)

Phase 1: modifying the tagset

Method (2/3)

Phase 1: modifying the tagset

1. Retagging of WSJ - introduction of determiner/pro-form distinction
2. Training Treetagger (Schmid 1994) on finer-grained tagset
3. Re-tagging learner and native corpora with the functional distinctions
   - DT just for determiner uses
   - TPRON for pronominal *this* or *that*
4. To be continued for other uses of *it* ...

Method (3/3)

Phase 2: Identifying forms

Method (3/3)

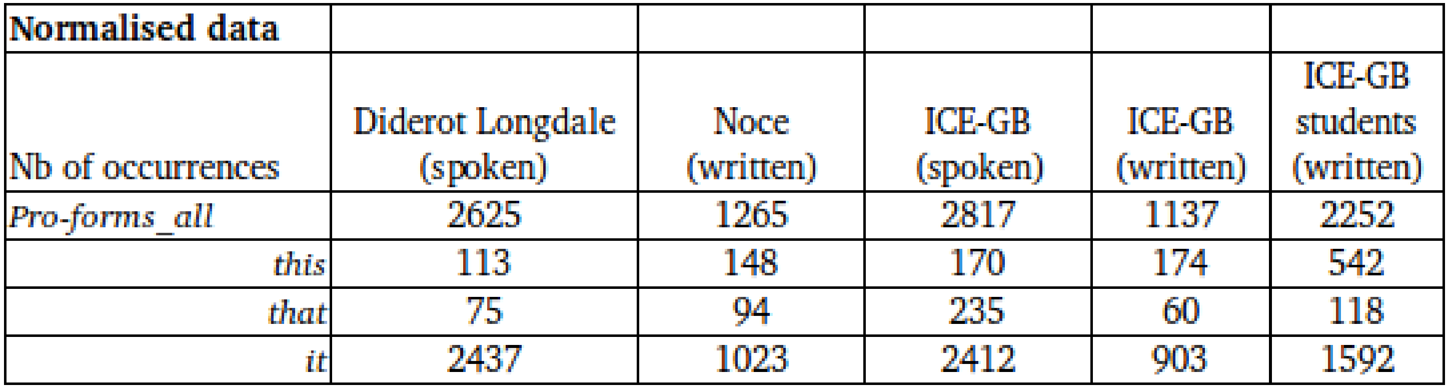
Phase 2: Identifying forms

- Function-form identification
  - Queries combine two layers: POS and text
  - NITE NXT (occurrence extractions) (Carletta et al. 2003)
    - ($u utterance)($wt word):$u^$wt & $wt@orth~/[tT]h(is)/ & $wt@pos="TPRON" ::($w word):$u^$w
  - AntConc (adjacency queries)
    - \b[t|T]his\b\tTPRON\n.*\t(V.*|MD)

Results and analysis (1/7)

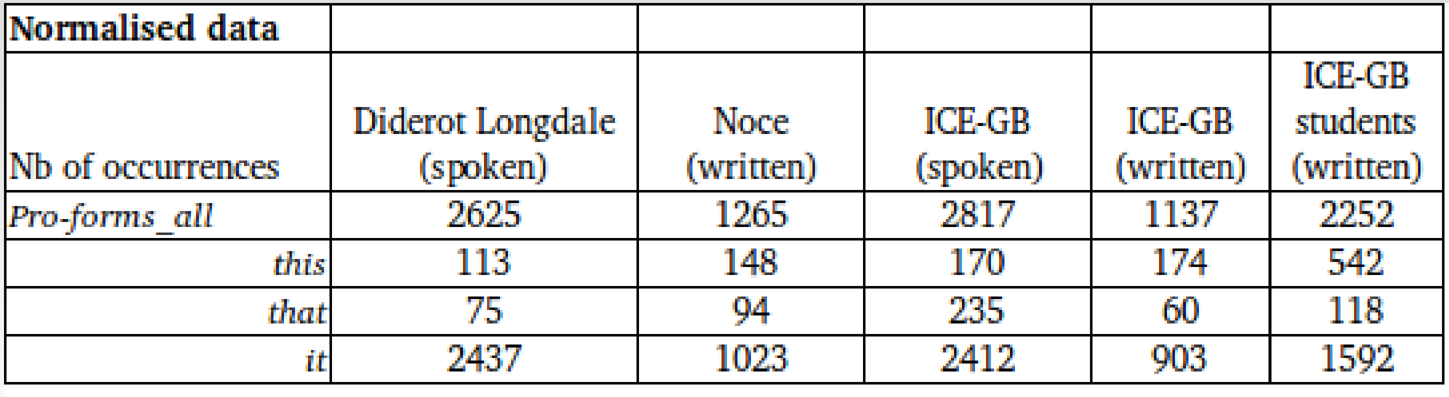
- Distributional study of the pro-forms and the pronoun *it*
- X^2 significant difference:

  X-squared = 768.3011, df = 8, p-value < 2.2e-16
- Caveats: sample validity - simple independence tests such as χ^2 not possible due to dependence of the data points (Gries, [2009] 2013:168)



Normalised data					
Nb of occurrences	Diderot Longdale (spoken)	Noce (written)	ICE-GB (spoken)	ICE-GB (written)	ICE-GB students (written)
Pro-forms_all	2625	1265	2817	1137	2252
this	113	148	170	174	542
that	75	94	235	60	118
it	2437	1023	2412	903	1592

Results and analysis (2/7)

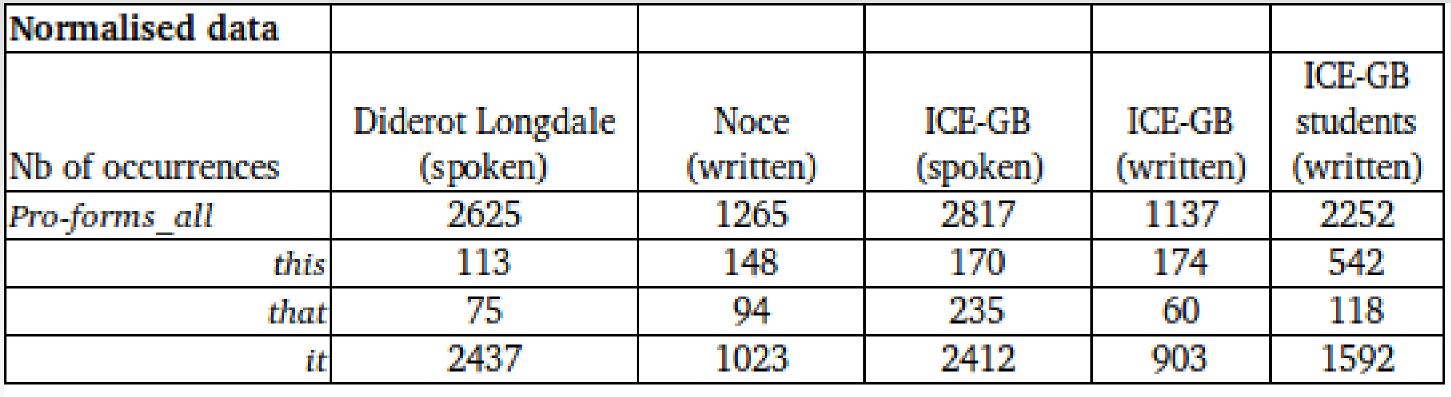


Normalised data					
Nb of occurrences	Diderot Longdale (spoken)	Noce (written)	ICE-GB (spoken)	ICE-GB (written)	ICE-GB students (written)
Pro-forms_all	2625	1265	2817	1137	2252
this	113	148	170	174	542
that	75	94	235	60	118
it	2437	1023	2412	903	1592

Pro-forms across corpora

- Use of *it* predominant in NS and NNS – similar with Biber's findings (1999: 347)
- Similar uses between NS and NNS
- Positive transfers (Ellis 1994) among learners for the pro-form function – equivalent (but not identical) systems in L1s
- Strong effect of register + students written essays may reflect a reluctance to repeat words

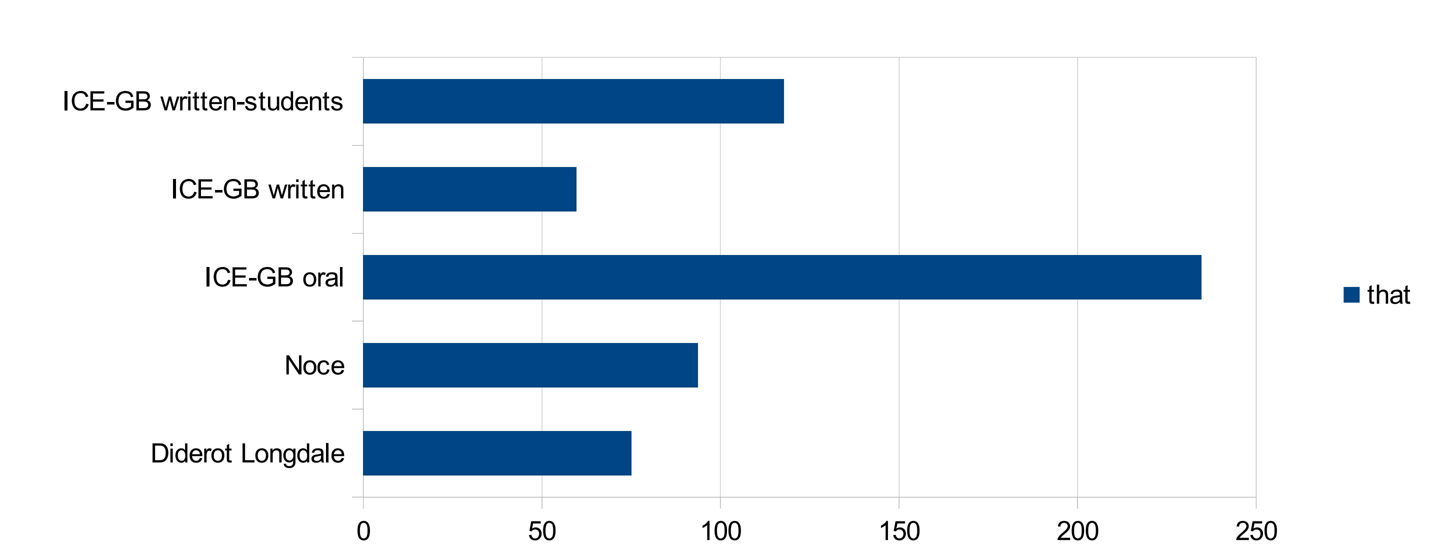
Results and analysis (3/7)



Normalised data					
Nb of occurrences	Diderot Longdale (spoken)	Noce (written)	ICE-GB (spoken)	ICE-GB (written)	ICE-GB students (written)
Pro-forms_all	2625	1265	2817	1137	2252
this	113	148	170	174	542
that	75	94	235	60	118
it	2437	1023	2412	903	1592

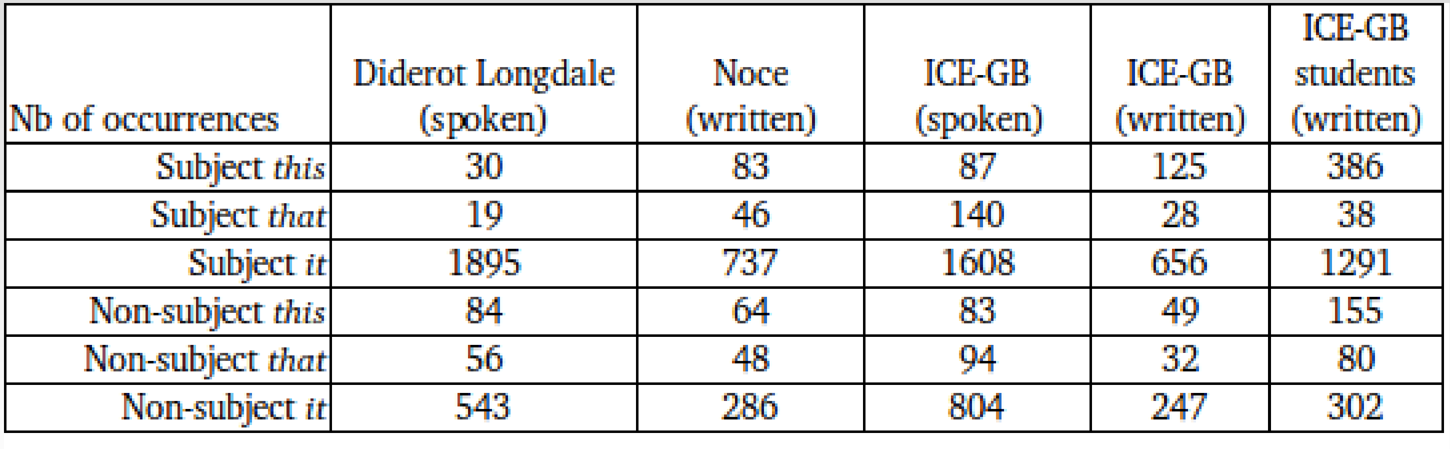
- Predominance of *it* → hidden trends for *this* and *that?*
- *This* is underused by learners

Results and analysis (4/7)



- Predominance of *it* → hidden trends
- *This* is underused by all learners
- *That* largely underused by learners of French L1 as opposed to other NNS → substitution strategy?

Results and analysis (5/7)

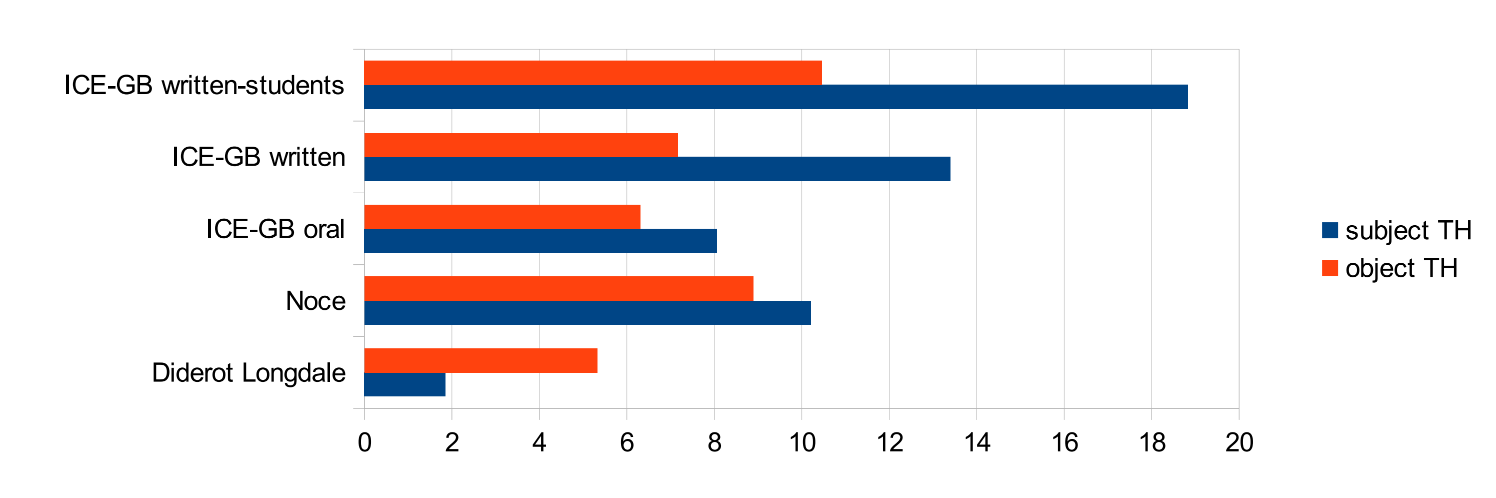


Nb of occurrences	Diderot Longdale (spoken)	Noce (written)	ICE-GB (spoken)	ICE-GB (written)	ICE-GB students (written)
Subject *this*	30	83	87	125	386
Subject *that*	19	46	140	28	38
Subject *it*	1895	737	1608	656	1291
Non-subject *this*	84	64	83	49	155
Non-subject *that*	56	48	94	32	80
Non-subject *it*	543	286	804	247	302

Pro-forms and their syntactic roles of subject

- Predominance of subject role (Precision and recall to be determined)
- Is it the same for all forms?

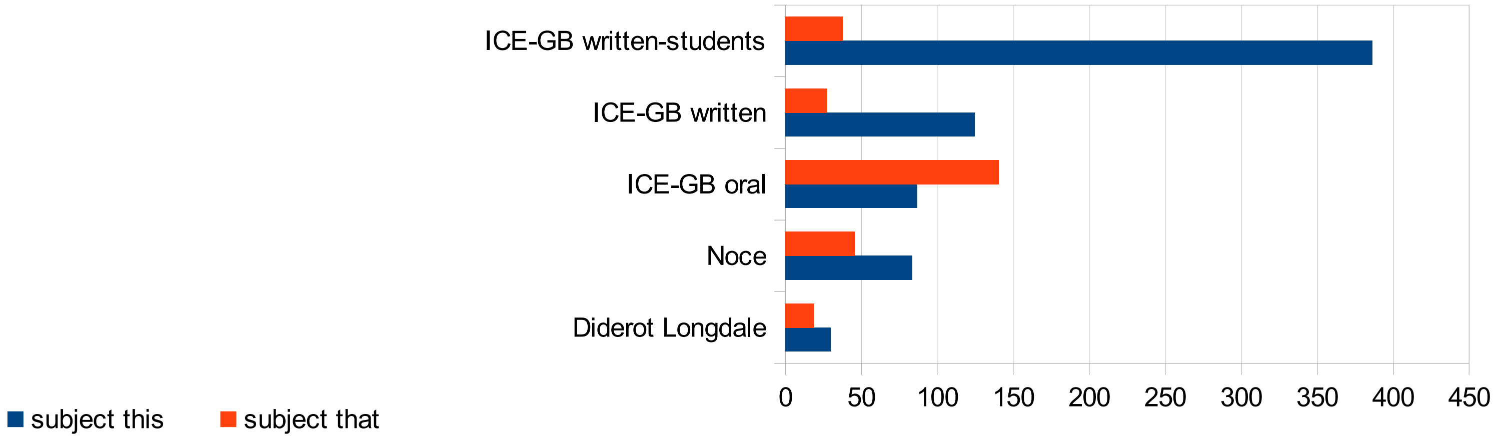
Results and analysis (6/7)



Pro-forms and their syntactic roles of subject

- Predominance of subject role along with Biber (1999:334)
- Is it the same for all forms?
- **Subject form <u>not</u> dominant for learners of French L1**

Results and analysis (7/7)



Pro-forms and their syntactic roles of subject

- Predominance of subject role
- Is it the same for all forms?
- Subject form **not** dominant for learners of French L1
- **Subject *that* even less and i*t* overused. So *it* the safe option?**

Example and hypothesis to be tested

The past tense triggers *that* for learners

“we haven't (er) during (er) twelve hours so for the food it wasn't very great and but (er) I want I didn't (er) (er) I wasn't ill so . that was <begin laughter> (er) nice <end laughter> (em) about (er) the traditions (er) it was like in (er) every every African culture traditions and it was interesting to know how this (er) . how it was and (em) .. there . there w= were (er) cyber cafes so <laughs> (er) for Internet ***that was*** nice to: . to give news by e-mails to our family” DID00066-S001

Hypothesis to be tested

Logistic regression of nested/inter-dependent variables

- CORPORA: REGISTER(*spoken/written*)
- L1(French/Spanish)
- THISTHAT(*this* or *that*)
- FUNCTIONAL REALISATION(Determiner/Pro-form)
- SYNTACTIC ROLE(Subject/Object)

Summary & outlook

- Work on PoS tags → functional approach with access to:
    - Information on facilitation and substitution strategies for pro-forms
    - Discourse analysis on information packaging subjecthood of pro-forms.
- Possible exploration of co-occurrences (*that* with past and *this* with present)
- Age as variable? In the case of student essays: interiorisation of the norm for NS (no repetitions and overuse of *this*)

Thanks

- To Ana Diaz Negrillo for sharing the NOCE corpus
- To Jonathan Kilgour for his help on NITE NXT queries

- thomas.gaillat@univ-paris-diderot.fr

References

- Barlow, Michael. 2005. « Computer-based analyses of learner corpora ». In Analysing Learner Language, by Rod Ellis et Gary Barkhuizen, 337-357. Oxford: Oxford University Press.
- Biber, Douglas, Stig Johanson, Geoffrey Leech, Susan Conrad, et Edward Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow: Longman.
- Cornish, Francis. 1999. Anaphora, Discourse, and Understanding. Evidence from English and French. Oxford: Oxford University Press.
- Diaz Negrillo, Ana. 2007. « A Fine-Grained Error Tagger for Learner Corpora ». Jaen.
- Ellis, Rod. 1994. The study of second language acquisition. 1 vol. Oxford applied linguistics 1995. Oxford, Royaume-Uni: Oxford university press.
- Ellis, Rod, et Gary Barkhuizen. 2005. Analysing learner language. Oxford: Oxford University Press.
- Fraser, Thomas, et André Joly. 1979. « Le système de la deixis - Esquisse d'une théorie d'expression en anglais ». Modèles linguistiques 1 (2): 97-157.
- ———. 1980. « Le système de la deixis (2) : Esquisse d'une théorie d'expression en anglais ». Modèles linguistiques 2 (2): 22-49.
- Gaillat, Thomas. 2013. « Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank ». In Proceedings of TALN13. http://www.taln2013.org/actes
- Gries, Stefan Thomas. 2009. Statistics for linguistics with R: a practical introduction. 1 vol. Trends in linguistics. Studies and monographs, ISSN 1861-4302 208. Berlin, Germany, United States.
- Halliday, M. A. K., et Ruqaiya Hasan. 1976. Cohesion in English. English Language Series. Harlow: Pearson Education Limited.
- Meunier, Fanny, Sylviane Granger, Damien Littré, et Magali Paquot. 2008. « The LONGDALE (Longitudinal Database of Learner English) ». UCL-CECL. http://www.uclouvain.be/en-cecl-longdale.html.
- Nelson, Gerald, Sean Wallis, et Bas Aarts. 1998. The British Component of the International Corpus of English (ICE-GB) and ICECUP software (CD-ROM) (version 3.1). London. http://www.ucl.ac.uk/english-usage/projects/ice-gb/.
- Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In Proceedings of the International Conference on New Methods in Language Processing, 14-16. Manchester: UK.

PoS tagging of corpora

WSJ	Rappel %	Précision %	F-Score %	True occurrences expected
This DT	100	91,04	95,31	61
This TPRON	60	100	75	15
That DT	75	78,94	76,94	20
That TPRON	55	88,23	68,18	27

Longdale	Rappel %	Précision %	F-Score %	True occurrences expected
This DT	93, 75	78,94	85,71	16
This TPRON	33,33	66,66	44,44	6
That DT	0	0	0	2
That TPRON	0	0	0	11