Towards a Diagnosis of Textual Difficulties for Children with Dyslexia

Solen Quiniou, Béatrice Daille
Université de Nantes, LS2N
2 rue de la Houssinière, 44322 Nantes Cedex 3, France
{Solen.Quiniou, Beatrice.Daille}@univ-nantes.fr

Abstract

Children's books are generally designed for children of a certain age group. For underage children or children with reading disorders, like dyslexia, there may be passages of the books that are difficult to understand. This can be due to words not known in the vocabulary of underage children, to words made of complex subparts (to pronounce, for example), or to the presence of anaphoras that have to be resolved by the children during the reading. In this paper, we present a study on diagnosing the difficulties appearing in French children's books. We are more particularly interested on the difficulties coming from pronouns that can disrupt the story comprehension for children with dyslexia and we focus on the subject pronouns *"il"* and *"elle"* (corresponding to the pronoun *"it"*). We automatically identify the pleonastic pronouns (*e.g.*, in *"it's raining"*) and the pronominal anaphoras, as well as the referents of the pronominal anaphoras. We also detect difficult anaphoras that are more likely to lead to miscomprehension from the children: this is the first step to diagnose the textual difficulties of children's books. We evaluate our approach on several French children's books that were manually annotated by a speech therapist. Our first results show that we are able to detect half of the difficult anaphorical pronouns.
Keywords: anaphora, dyslexia, children's book, French

1. Introduction

The democratization of books on digital tablets has allowed the design of new methods to support people with reading troubles. Children's book publishers have thus proposed adaptations for young readers, with the use of specific typefaces, larger margins and text spaces, or refined illustrations to help them to better understand the content of the text. These adaptations have also proven their efficiency on Web browsers as they have allowed readers with dyslexia to better use their short-term memory on the text and on its meaning (Parilova et al., 2016). Research works have been mainly focused on using audio interfaces, as speech dictation, and screen readers to offer new features in digital books addressed to readers with dyslexia (Sitbon et al., 2007) rather than on supporting the difficulties coming from the content of the text itself.

Our work is complementary of the previous approaches as it only considers the content of the text. Indeed, we are interested in evaluating the difficulties that occur in a text, in the form of textual ambiguities, and that can lead to difficulties in the comprehension of the story, especially for children with reading troubles like dyslexia. There are ambiguities at various levels: at the phonetic level such as the French word *"fils"* that can be translated as *"son"* or *"threads"* depending on the context and which is either a singular noun, or a plural noun (furthermore, its pronunciation is different in both cases); at the lexical level, *e.g.* due to collocations like *"il pleut des cordes"* that translates into *"it's raining cats and dogs"* (whereas *"cordes"* usually translates into *"ropes"*); at the pragmatical level as in *"il lui parle"* (*"he talks to him/her"*) where *"il"* is an anaphorical pronoun which referent has to be found between the preceding masculine nominal groups and *"lui"* is also an anaphorical pronoun which referent can be either a masculine, or a feminine nominal group.

To our knowledge, in the field of natural language processing, the difficulties coming from dyslexia have been studied in the writing of children with dyslexia (Rello et al., 2016; Rauschenberger et al., 2016) but not in their reading. Evaluating the difficulty of a text is a field of natural language processing that has been recently studied (François and Watrin, 2011; Gala et al., 2014; Ho Dac et al., 2016; Tanguy et al., 2016; Müller et al., 2016). Nonetheless, these works rather focused on predicting the lexical complexity and the graduation of words in a text whereas we are more interested in establishing a diagnosis of the difficulties appearing in a text but also to propose an explanation for the identified difficulties. Indeed, our goal is to allow the visualization of the difficulties on digital children's books and also to be able to visualize the explanations of the detected difficulties.

2. Reading Difficulties of Children with Dyslexia

According to the French dictionary of speech therapy (Brin-Henry et al., 2011), dyslexia is " the term used to name all the specific and durable troubles that manifest themselves when a person (child or adult) identifies written words during reading. Dyslexic troubles persist throughout the life course of the person. The World Health Organization estimates that 8 % to 12 % of the world's population is affected by these dyslexic troubles. "

In practice, dyslexia can express itself in various ways: omission, substitution, sound inversion in words; confusion between mirror letters (*"d/b"* and *"p/q"*) and between close sounds (*"ch/j"* and *"d/t"*); decoding difficulty; guessing words relying on their first letters or on the meaning of the sentence; difficulties to recog-

Book title	Code	Age group	#Words	#Personal pronouns	#*"il"*	#*"elle"*
Ali Baba et les quarante voleurs	ALI	8-12	2 458	258	80	29
L'arbre et le bûcheron	ARB	8	1 517	161	42	2
Le buveur d'encre	BUV	7	1 002	161	34	2
Dans le ventre du cheval de Troie	CHE	9	2 623	266	50	11
Emporté par le vent	EPV	12-15	3 005	253	30	142
Nos étoiles contraires	NEC	14	35 484	3 055	595	234

Table 1: Presentation of the children's books in the corpus

nize words (small lexical stock); difficulties for irregular words; skipping words or lines; comprehension difficulties due to the attention focused on the decoding; misreading a word for another one (paralexia).

In this work, we focus on ambiguities caused by pronouns, and especially anaphoras (Botley and McEnery, 2000). For children with dyslexia, they can lead to comprehension difficulties in several situations:

- in the French language, the pronoun *"il"* can be either pleonastic, or anaphorical. When the pronoun is pleonastic, if the children do not detect it, they will be looking for a referent that does not exist;

- the pronouns *"lui"* and *"leur(s)"* (corresponding to *"her/him/it"* and *"their"*) can be either feminine, or masculine. For example, in the sentence *"Il lui(leur) laissa la vie sauve"* (*"He spared his/her(their) life(s)"*), we do not know if the referent of *"lui(leur)"* is feminine or masculine;

- when the referent is located several sentences before the pronoun, the children will have trouble to find it. For example, in ALI (see Section 3.), we have the following translated sentences: *"Ali Baba is happy for his brother. Good thing that Cassim is married to a rich heir! Good thing that he is an important merchant of the city! But too bad that Cassim does not want to see him anymore."*. In the last sentence, *"him"* refers to *"Ali Baba"* but it appears 3 sentences before this pronoun;

- in dialogues, the pronouns *"je"* and *"tu"* (*"I"* and *"you"*) do not always refer to the same person as it depends on the person who is talking. It makes it hard for children with dyslexia to follow the dialogue, especially in the case of long dialogues;

- words like *"en"* or *"le/la/les"* can be either pronouns, or prepositions and determiners: it makes it hard to understand them. For example, in *"il se déplaçait en silence"* (*"he was moving silently"*), *"en"* is a preposition, whereas in *"l'eau coule ; il y en a partout"* (*"water flows; it is everywhere"*), *"en"* is a pronoun;

- when there is an anaphorical chain, *i.e.* a sequence of pronouns corresponding to the same referent, it might be helpful to annotate the first pronoun because it is more difficult to retrieve than the following ones. For example, in ALI, we have the following translated anaphorical chain: *"Still trembling, Ali Baba comes down the tree. He knows that he must quickly leave this place! He should not get involved in this! This is too dangerous! He is only a modest logger!"*

- when the referent is located after the pronoun (*i.e.* in a cataphora), finding the referent will be harder for the children. For example, in ALI, we have the following translated cataphora: *"Cassim's wife goes in the kitchen to pick up the jar herself. When she gets back home, Ali Baba's wife plunges the jar in the gold coins."* Here, *"she"* refers to *"Ali Baba's wife"* and not to *"Cassim's wife"*;

- when several clitics pronouns appear between the subject pronoun and the verb, it is difficult to identify the referent of each pronoun. For example, in ARB, we have *"je te l'ai dit"* (*"I told it to you"*) where *"te"* and *"l'"* are two pronouns.

3. Corpus

The raw corpus is first described and the manual annotation process is then presented.

3.1. Corpus Used in this Study

The corpus is provided by Mobidys[1], a startup that publishes digital books for children with dyslexia. The corpus gathers full-texts or extracts of French children's books (some of them have been adapted to children with dyslexia by rewriting the text of the books). Table 1 gives the corpus features: the age group to which the books are addressed, the number of words, the number of personal pronouns, and the number of *"il"* and *"elle"* pronouns. We can see that the proportion of personal pronouns is different between the books and goes from 8.4% (for EPV) to 16.1% (for BUV): books addressed to younger children contain a higher proportion of personal pronouns as compared

[1] www.mobidys.fr

to books addressed to teenagers. This gives an indication on the textual difficulty: a text with a higher proportion of pronouns is more likely to contain a higher number of ambiguities and thus to be more difficult to understand for children with dyslexia. Furthermore, *"il"* pronouns represent 11.9% to 31.0% of the personal pronouns whereas *"elle"* pronouns represent 1.2% to 56.1% of them: they thus represent about one third of the personal pronouns in the corpus.

3.2. Corpus Annotation

To evaluate the detection of pleonastic pronouns and anaphoras, a speech therapist student (in her fourth year of study) annotated the occurrences of *"il"* pleonastic pronouns, and of *"il"* and *"elle"* pronominal anaphoras with their referents. She also judged the difficulty of the pronominal anaphoras. Table 2 summarizes these numbers for each book[2] (the rate of difficult anaphoras is computed according to the number of anaphorical pronouns).

Text	#Pleonastic pronouns	#Anaph. pronouns	#Difficult anaphoras
ALI	7	102	10 (9.8%)
ARB	9	35	9 (25.7%)
BUV	10	26	7 (26.9%)
CHE	20	41	2 (4.9%)
EPV	8	164	18 (11.0%)
NEC	17	85	7 (8.2%)

Table 2: Number of manual annotations of *"il"* and *"elle"* pronouns in the corpus

Anaphorical pronouns represent the majority of the *"il"* and *"elle"* pronouns (only *"il"* pronouns can be pleonastic). In ARB and BUV, a quarter of the anaphoras were considered difficult whereas the books are addressed to young children (7-8 year olds).

4. Pronoun Detection and Resolution

Difficult anaphora detection consists in three main steps: the pleonastic pronoun detection, the anaphora resolution and, the difficult anaphora diagnosis.

4.1. Pleonastic Pronoun Detection

In the French language, the *"il"* pronoun is either a subject pronoun, or a pleonastic pronoun. As the anaphora resolution only applies to argument pronouns, it is necessary to distinguish between the two types of *"il"* pronouns. To do so, we define rules over sequences of tokens with their grammatical features, each rule being in the form of a regular expression. Specific components were developed to deal with regular expressions on annotations, such as TokenRegex(Chang and Manning, 2014). We use PyRATA[3], available for Python, which uses part-of-speeches, inflectional forms, and lemmas in the regular expressions. From the set of rules written by Danlos (2005) to disambiguate the *"il"* pronouns, we kept 15 rules: those exclusively describing a pleonastic use, those including typical collocations, and a subpart of those describing current pleonastic sentence structures with an extraposed nominal subject. Here are examples of rules, for each category:

```
(1) word="il" word="ne"? lemme@"meteo"
(2) word="il" lemme="ne"? lemme="y"
    lemme="en"? lemme="avoir"
(3) word="il" word="ne"? pos="PRO:PER"?
    lemme="avoir"? word="pas"? lemme="manquer"
```

Rule (1) recognizes pleonastic sentences with weather verbs belonging to the meteorological semantic class, such as *"il neige"* (*"it snows"*). Rule (2) describes the typical collocation *"il y a"* (*"there is"*). Rule (3) expresses the pleonastic sentence structure where the subject of the verb *"manquer"* (*"to lack"*) occurs as a direct object and the subject is the pleonastic pronoun, *"il manque du pain"* (*"there is a bread shortage"*). We do not take into account Danlos rules belonging to the formal language register as it does not characterize children's book. Furthermore, rule (3) is ambiguous because it can miscategorize the *"il"* pronoun as pleonastic when it is anaphorical. Indeed, in *"il manque une évaluation "* (*"an evaluation is missing"*), *"il"* is pleonastic, but in *"il manque de confiance en lui"* (*"he lacks self-confidence"*), *"il"* is anaphorical. We tag these ambiguous structures as UNDETERMINED as they are likely to confuse children with dyslexia.

4.2. Anaphora Resolution

To perform pronominal anaphora resolution, we apply the knowledge-poor approach of Mitkov (2002): it only requires part-of speech tagging and chunk identification as the linguistic preprocessing. The algorithm identifies the nominal chunks that precede an anaphorical pronoun, within a distance of two sentences, then checks the inflectional agreement with the anaphora and finally applies indicators to rank the nominal chunks. Each indicator gives either a positive or a negative score. The nominal chunk with the highest combined score is chosen as the antecedent of the anaphorical pronoun. We use RDRPOSTagger (Nguyen et al., 2014) to extract the nominal chunks. This tagger is designed for French and gives the gender and the number of a word: these informations are used for the inflectional agreement part of the anaphora resolution.
Mitkov (2002) listed 10 indicators. We kept 6 indicators as such: definiteness, givenness, lexical reiteration, non prepositional noun phrases, collocation pattern preference, and referential distance. We adapted two other indicators, *i.e.* section heading preference

[2]Due to a lack of time, only the first pronouns of NEC were annotated. The annotation of the remaining pronouns is in progress.

[3]https://github.com/nicolashernandez/PyRATA

and domain terminology preference, to the type of our texts. To represent the section heading preference, we only used the book title as there is no section. To take into account the domain terminology preference, we used the list of characters of the book. This list has to be manually supplied. We assign a higher score to main characters than to secondary or peripheral characters. We removed two indicators, indicating verbs and immediate reference, as they did not fit with linguistic features of children's books. We apply the eight indicators on each nominal chunks and we only keep the fourth first chunks with a combined score greater than two. When no antecedent is found, we assign the antecedents of the previous anaphorical pronoun to this anaphorical pronoun.

4.3. Difficult Anaphora Diagnosis

We consider as difficult anaphoras, the following anaphorical pronouns:

- pronouns with a distance of more than two sentences with the antecedent;
- pronouns with a high number of antecedents, *i.e.* three ore more;
- pronouns at the start of an anaphorical chain;
- pronouns belonging to an anaphorical chain of five pronouns or more;
- pronouns with several character names among the antecedents.

Nonetheless, difficulties due to anaphorical chains were not always tagged as such in the reference corpus. For example, in ALI, there is an anaphorical chain of seven anaphorical pronouns but none of them were considered to be difficult.

5. Experimental Results

The results of our experiments on the 3 parts of the pronoun detection and resolution are given in Table 3 with respect to the manual annotation of the corpora. They are discussed in the following subsections.

Text	Impers. pronouns		Resolved pronouns		Difficult anaphoras	
	REC	**PR**	**REC**	**PR**	**REC**	**PR**
ALI	85.7%	100%	88.2%	89.1%	60.0%	16.7%
ARB	77.8%	100%	74.3%	74.3%	33.3%	17.6%
BUV	90.0%	100%	73.1%	73.1%	42.9%	60.0%
CHE	90.0%	100%	68.3%	68.3%	50.0%	5.0%
EPV	25.0%	100%	72.6%	71.3%	27.8%	23.8%
NEC	52.9%	100%	49.4%	47.2%	42.9%	6.0%

Table 3: Recall (REC) and precision (PR) on the pronoun detection and resolution in the corpus

5.1. Pleonastic *"il"* Detection

For all the texts, the precision reaches 100%. For the first four texts, the recall can reach 100% if we consider the UNDETERMINED pronouns to be pleonastic ones. For EPV, the recall goes to 62.5% by adding the UNDETERMINED pronouns. The remaining pleonastic pronouns correspond to pronouns used with the verbs *"pouvoir"* (*"can"*) and *"être"* (*"be"*). These verbs are more often used with anaphorical pronouns than with pleonastic ones. Furthermore, the negative form does not appear in the regular expressions: considering it would allow an additional increase in the recall.

5.2. Anaphora Resolution

First results are quite satisfactory, except for NEC. This is mainly due to anaphorical chains because if the referent of the first pronoun is wrong, the mistake will be propagated to the other pronouns of the anaphorical chain. Another source of mistakes occur when the referent is not present in one of the two preceding sentences but in a sentence before these two sentences. The last source of mistakes comes from the POS tagger which wrongly tags some chunks as nominal ones and thus allows them to be candidate referents. For example, in NEC, 20% of the anaphorical *"il"* referents are not nominal chunks: it includes *"ai pardonné"* (*"has forgiven"*), *"allume"* (*"turn on"*), or *"musclé"* (*"strong"*).

5.3. Difficult Anaphora Diagnosis

The recall of difficult anaphoras is better than the precision which is quite low. Our approach tends to over detect difficult anaphoras. It is also due, in part, to the POS tagger and the misdetection of nominal chunks (or the misdetection of the frontiers of the chunks). Indeed, each time a pronoun corresponds to an unseen nominal chunk, this pronoun is considered to be difficult. For example, in ARB, *"l'arbre"* (*"the tree"*) and *"à l'arbre"* (*"to the tree"*) are 2 detected nominal chunks: the second one should be just *"l'arbre"* and not a new referent that would be also further detected as a difficult anaphora.

6. Conclusion

In this paper, we have presented a first attempt to diagnose difficulties in children's books. Our work focused on distinguishing pleonastic pronouns from anaphorical pronouns, and on recovering the referents of anaphorical pronouns for *"il"* and *"elle"* pronouns (corresponding to the *"it"* pronoun). We also proposed a first attempt at identifying difficult anaphoras, *i.e.* anaphoras that are more likely to cause difficulties in the comprehension of children with dyslexia.

Currently, we are designing an experimental evaluation on the detection of difficult anaphoras with children suffering from dyslexia as well as with other children. In future works, we want to extend the diagnosis to other pronouns as well as to the vocabulary used in the text (as compared to children age groups) and to the detection of words with complex subparts to read.

7. Acknowledgements

This work is part of the AmbiDYS project which has been funded by the RFI OIC (Région Pays de la Loire). We would also like to thank Kevin Espasa and David Kerbrat who worked on this project as interns as well as the Mobidys startup which provided the children's books we worked on.

8. Bibliographical References

Botley, S. P. and McEnery, T. (2000). *Corpus-based and Computational Approaches to Discourse Anaphora*. John Benjamins Publishing Company.

Brin-Henry, F., Courrier, C., Lederlé, E., and Masy, V. (2011). *Dictionnaire d'orthophnie - 3e édition*. Ortho Edition.

Chang, A. X. and Manning, C. D. (2014). TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Department of Computer Science, Stanford University.

Danlos, L. (2005). Ilimp: Outil pour repérer les occurrences du pronom impersonnel il. In *Actes de TALN*, pages 123–132.

François, T. and Watrin, P. (2011). Quel apport des unités polylexicales dans une formule de lisibilité pour le français langue étrangère ?

Gala, N., François, T., Bernhard, D., and Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots.

Ho Dac, L. M., Muller, S., and Delbar, V. (2016). L'anticorrecteur : outil d'évaluation positive de l'orthographe et de la grammaire.

Mitkov, R. (2002). *Anaphora resolution*. Longman.

Müller, A., François, T., Roekhaut, S., and Fairon, C. (2016). Classification automatique de dictées selon leur niveau de difficulté de compréhension et orthographique.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. (2014). RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In *Proceedings of the Demonstrations at EACL*, pages 17–20.

Parilova, T., Mrvan, B., Mizi, B., and Hldka, E. (2016). Emerging technology enabling dyslexia users to read and perceive written text correctly.

Rauschenberger, M., Rello, L., Füchsel, S., and Thomaschewski, J. (2016). A language resource of german errors written by children with dyslexia. In *Proc. of LREC*.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2016). DysList: An annotated resource of dyslexic errors. In *Proc. of LREC*.

Sitbon, L., Bellot, P., and Blache, P. (2007). éléments pour adapter les systèmes de recherche d'information aux dyslexiques. *Revue TAL*, 48(3):1–26.

Tanguy, L., Fabre, C., and Mercier, C. (2016). Analyse d'une tâche de substitution lexicale : quelles sont les sources de difficultés ?