Traitement des similarités dans la justification de paragraphe

Exposé GUTenberg, 10 septembre 2025

Didier Verna

didier@didierverna.net

didierverna.net @didierverna didier.verna in/didierverna

Préambule

- Article à ACM SIGWEB DocEng 2024

Similarity Problems in Paragraph Justification

An Extension to the Knuth-Plass Algorithm

Didier Verna
EPITA Research Laboratory
Le Kremlin-Bicêtre, France
didier@lrde.epita.fr

Abstract

In high quality typography, consecutive lines beginning or ending with the same word or sequence of characters is considered a defect. We have implemented an extension to TeX's paragraph justification algorithm which handles this problem. Experimentation shows that getting rid of similarities is both worth addressing and achievable. Our extension automates the detection and avoidance of similarities while leaving the ultimate decision to the professional typographer, thanks to a new adjustable cursor. The extension is simple and lightweight, making it a useful addition to production engines.

CCS Concepts

• **Applied computing** → **Document preparation**; • **Theory of computation** → *Dynamic graph algorithms*.

Keywords

Paragraph Justification, Similarity Avoidance, Homeoteleutons, Homeoarchies, TeX, Knuth-Plass Extension

ACM Reference Format:
Didier Verna. 2024. Similarity Problems in Paragraph Justification: An Extension to the Knuth-Plass Algorithm. In *ACM Symposium on Document Engineering 2024 (DocEng '24), August 20–23, 2024, San Jose, CA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3685650.3685666

1 Introduction

In spite of its relatively old age, Donald Knuth's TeX typesetting system [10, 11] is still considered a *de facto* standard when it comes to digital typography. In particular, its paragraph justification algorithm, known as the *Knuth-Plass* [12], established a landmark in the category of algorithms considering a paragraph as a whole rather than proceeding line by line, as earlier (greedy) algorithms used to do [1, 4, 16].

Yet, many aspects of fine typography are not directly or automatically handled by TeX. Consider for example the leftmost paragraph in Figure 1. The typesetting was done by TeX with the Latin Modern Roman font at a 10pt size, and for a paragraph width of 201pt (the figure is optically scaled down in order to fit on the page).

Notice how three lines near the end of the paragraph end the same way, with the word "and". This is considered a defect in high quality typography, as it generates a micro-interruption: the reader's attention may be caught by the similarity and temporarily loose focus or concentration. Such similarities may also lead the reader to accidentally skip a line or re-read the same one, even more so when the problem occurs at the beginning of the line rather than at the end of it. In fact, the field of textual criticism has identified the problem and its consequences in the very ancient context of scribal errors (*e.g.* missing lines in manual copies of the bible made by monks) [20].

We have implemented an extension to the Knuth-Plass (KP) algorithm that is able to deal with that kind of defect. In this paper, we use the term *similarity* for the lack of an official terminology. We would also like to propose the expression *character / word ladder*, analogous to the more widespread expression *hyphenation ladder*, the meaning of which should be obvious. Accidental line skipping has been referred to with a rather awkward French expression, *saut du même au même* ("jump from same to same"), even in non French litterature [20]. Borrowed from rhetoric, the terms *homeoarchy* and *homeoteleuton* have come to designate beginning and end of line similarities respectively, and by extension, accidental line skipping because of them [15].

This paper is organized as follows. Section 2 mentions some related work. Section 3 provides an outline of the KP algorithm's operation, necessary to understand how our extension works. Section 4 describes our extension, and Section 5 presents some experimental results.

2 Related Work

Frank Mittelbach mentions the similarity problem in a survey of existing alternative TeX engines and remaining issues [13]. No solution is proposed in this paper, as it is merely a state-of-the-art review. Alex Holkner addresses the problem in his multiple-objective approach to line breaking [7]. However, the paper only seems to be concerned with beginning-of-line similarities (called *stacks*). Although inspired from it, the approach is not technically an extension to TeX's algorithm. The paper does not provide a precise definition for stacks, and does not use the corresponding objective function in the reported experimental results. Other extensions to the KP algorithm have been proposed in the past, some micro-typographic [17], some macro-typographic, for example to help with automatic document layout [6]. Concerning the latter, our underlying motivations are in fact opposite. The work in question attempts to provide flexibility *at the expense of quality* in order to cope with situations in which manual intervention is impossible, such as automatically adjusting to different displays. We, on the other hand, are interested

DocEng '24, August 20–23, 2024, San Jose, CA, USA

ACM ISBN 979-8-4007-1169-5/24/08
https://doi.org/10.1145/3685650.3685666

https://didierverna.net/publications/verna.24.doceng/

Similarités

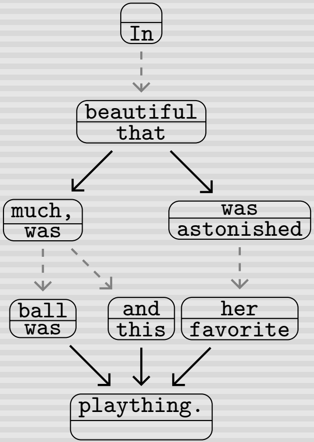
In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

- Solution de TeX
  - Largeur : 201 pt
  - Caractère : Latin Modern Roman 10 pt
- Problèmes (également en début de ligne)
  - Micro-interruption
  - Saut de ligne accidentel
- Vieille question (moines scribes)
  - Cf. philologie/analyse de texte
- Terminologie
  - « Saut du même au même » 🤢
  - « Cascade » (Thomas Savary)
  - 💡 Homéoarchie/Homéotéleute (rhétorique)
  - 💡 Échelle de caractères/mots

Solutions

- 👊 Espaces insécables (cf. exposé de Thomas Savary « Composer comme un pro »)
  - Homéoarchie : `~truc ... truc~`
  - Homéotéleute : `truc~ ... ~truc`
- 💪 Un peu (plus) de douceur (dans ce mo…)
  - Homéoarchie : `\penalty`xxx `truc ... truc\penalty`xxx
  - Homéotéleute : `truc\penalty`xxx `... \penalty`xxx `truc`
- Dans tous les cas
  - Travail manuel
  - $xxx = ??$
  - Non automatisable
- Automatiser !

Le Knuth-Plass en 30 secondes



- Problème de type « plus court chemin »
  Trouver la « meilleure » route du début jusqu'à la fin du paragraphe
- Technique d'optimisation dynamique
  Ne jamais construire l'intégralité du graphe
- Fonction de coût
  - Démérites locaux : ligne par ligne
    Badness, pénalité de césure, etc.
  - Démérites contextuels : ligne contre ligne
    Échelles de césure, disparités d'espacement, etc.
- Décision finale : $\min \sum (\textit{d. locaux} + \textit{contextuels})$

Extension au Knuth-Plass

- **Idée :** un nouveau type de démérites contextuels (« démérites de similarité »)
  - Comparer les débuts & fins de lignes consécutives
  - Ajouter les démérites de similarité le cas échéant
  - Note : cas particulier des deux lignes finales
- **Implémentation :**
  - Sauvegarder les séquences de début/fin de ligne dans les nœuds
  - Seulement jusqu'à la première « glue » (espace élastique)/premier « discrétionnaire » (point de césure/ligature)
  - Supprimer les « kerns »
- **Justification :**
  - Discrétionnaires : éviter le surcoût de la déconstruction 🤔
  - Glue : éviter les considérations d'alignement vertical
  - Kerns : petits ajustements, identiques en cas de similarité

Extension au Knuth-Plass

- **Idée :** un nouveau type de démérites contextuels (« démérites de similarité »)
  - Comparer les débuts & fins de lignes consécutives
  - Ajouter les démérites de similarité le cas échéant
  - Note : cas particulier des deux lignes finales
- **Implémentation :**
  - Sauvegarder les séquences de début/fin de ligne dans les nœuds
  - Seu… …nnaire » (point de c…
  - Sup…

Comparer des séquences courtes

Qui peut le moins peut le plus…

- **Justification :**
  - Discrétionnaires : éviter le surcoût de la déconstruction 🤔
  - Glue : éviter les considérations d'alignement vertical
  - Kerns : petits ajustements, identiques en cas de similarité

Démonstration

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits=0`

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits=2800`

In olden times when wishing still helped one, there lived a king whose daughters were all beautiful; and the youngest was so beautiful that the sun itself, which has seen so much, was astonished whenever it shone in her face. Close by the king's castle lay a great dark forest, and under an old lime-tree in the forest was a well, and when the day was very warm, the king's child went out into the forest and sat down by the side of the cool fountain; and when she was bored she took a golden ball, and threw it up on high and caught it; and this ball was her favorite plaything.

`\similardemerits=5230`

Valeur ajoutée

- **Pertinence :** le problème est-il fréquent ?
- **Efficacité :** cette extension le résout-il ?
- **Deux expériences orthogonales :**
  - Un paragraphe à des largeurs très variées
  - De nombreux paragraphes à une largeur fixe

Conditions expérimentales

- *Frog King* (des Frères Grimm) §1
- 142 pt ($\approx$ 5 cm) $\rightarrow$ 569 pt ($\approx$ 20 cm)
- 427 cas

Conditions expérimentales

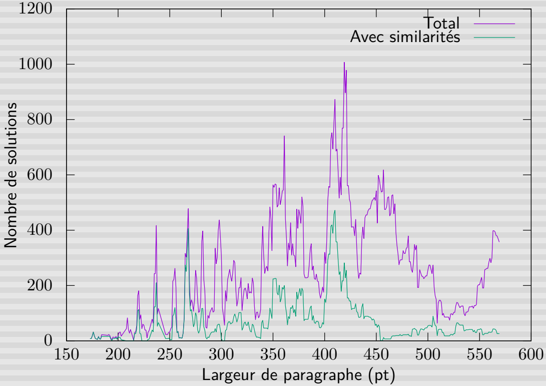
- Moby Dick (Herman Melville)
- 1 524 paragraphes à 284 pt ($\approx$ 10 cm)

- Total :
  - Expériences 1 & 2 $\rightarrow$ 1 951 cas
  - x 3 $\rightarrow$ 5 853 passes

 10 septembre 2025

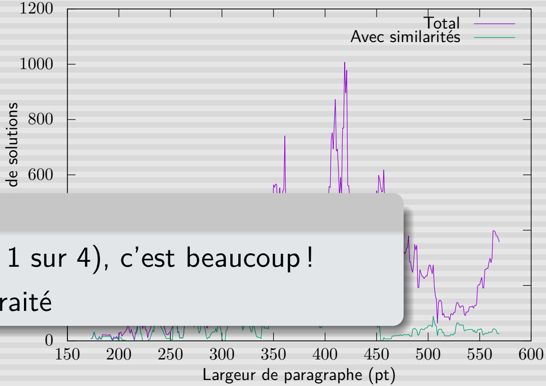
Pertinence (*Frog King*)

- Résultats attendus
  - Formes chaotiques
  - Valeurs proches pour des paragraphes étroits
- Découvertes
  - Similarités fréquentes
  - Solutions alternatives également
- Comportement de TeX
  - Similarités inévitables : 4 %
  - Solution contenant des similarités
    - Expérience 1 : 21 %
    - Expérience 2 : 26 %



Pertinence (*Frog King*)

- Résultats attendus
  - Formes chaotiques
  - Valeurs proches pour des paragraphes étroits
- Découvertes
  - Sim
  - Solu
- Comport
  - Sim
  - Solution contenant des similarités
    - Expérience 1 : 21 %
    - Expérience 2 : 26 %



Conclusion

- 2 paragraphes sur 10 (voire 1 sur 4), c'est beaucoup !
- Le problème mérite d'être traité

 10 septembre 2025

Efficacité

`\similardemerits=10000`	$+$ `\adjacentdemerits=0`
Corrigés : 48 %/50 % Améliorés : 50 %/63 %	Corrigés : 53 %/66 % Améliorés : 57 %/73 %

- Conclusion : Traitement automatique possible
- ⚠ Attention !
  - Colonne de droite = conditions extrêmes
  - Moins de similarités ne signifie pas forcément meilleure esthétique
  - Mais il y a une marge de manœuvre

Conclusion

- Traitement des similarités pour le Knuth-Plass
  - Extension simple & légère
  - Rétrocompatible avec TeX (`\similardemerits=0`)
- Implémenté dans ETAP, utile dans les moteurs en production
- Expérimentation → traitement automatique des similarités possible et souhaitable

Perspectives

- Étude de l'interaction avec les autres critères esthétiques
- Démérites de similarité en tant que fonction plutôt que valeur scalaire
- LuaMetaTeX ⟶ `\left/right twindemerits`

