Computational Linguistics in the Netherlands Journal 4 (2014) 53-74 Submitted 05/2014; Published 12/2014

T-Scan: a new tool for analyzing Dutch text

Henk Pander Maat[*] H.L.W.PANDERMAAT@UU.NL
Rogier Kraf[*] KRAF@ZIGGO.NL
Antal van den Bosch[**] A.VANDENBOSCH@LET.RU.NL
Nick Dekker[*] N.W.DEKKER@UU.NL
Maarten van Gompel[**] PROYCON@ANAPROY.NL
Suzanne Kleijn[*] S.KLEIJN1@UU.NL
Ted Sanders [*] T.J.M.SANDERS@UU.NL
Ko van der Sloot[***] KO.VANDERSLOOT@UVT.NL

[*] *Department of Languages, Literature and Communication, Utrecht University*

[**] *Centre for Language Studies, Radboud University Nijmegen*

[**] *Department of Communication and Information Sciences, Tilburg School of Humanities*

Abstract

T-Scan is a new tool for analyzing Dutch text. It aims at extracting text features that are theoretically interesting, in that they relate to genre and text complexity, as well as practically interesting, in that they enable users and text producers to make text-specific diagnoses. T-Scan derives it features from tools such as Frog and Alpino, and resources such as SoNaR, SUBTLEX-NL and Referentie Bestand Nederlands.

This paper offers a qualitative discussion of a number of T-Scan features, based on a minimal demonstration corpus of six texts, three of them scientific articles and three of them drawn from a women's magazine. We discuss features concerning lexical complexity, sentence complexity, referential cohesion and lexical diversity, lexical semantics and personal style. For all these domains we examine the construct validity as well as the reliability of a number of important features. We conclude that T-Scan offers a number of promising lexical and syntactic features, while the interpretation of referential cohesion/ lexical diversity features and personal style features is less clear. Further developing the application and analyzing authentic text need to go hand in hand.

1. Introduction

T-Scan is a software tool for analyzing Dutch text, available for researchers at webservices-lst-science-ru.nl. In its present state, it returns around 300 features for a text; these features may be aggregated to the text level, but may also be examined per sentence and per word. The tool is currently used by a small (but growing) number of text and discourse researchers, who use it to compare texts on a number of stylistic indices. In the near future, we will use the tool in readability prediction and genre classification studies. While such studies may primarily focus on prediction success, we intend to relate our work to linguistic text typology (e.g. Biber & Conrad 2009) and text and language processing research. In the long run, we want T-Scan to provide professional writers with feedback on texts: T-Scan should eventually tell them something about the complexity and the genre-conformity of their texts and point out what aspects of the text may need reconsideration.

The automatic analysis of text difficulty started with the traditional readability work by researchers such as Bormuth (1966), Klare (1963) and Dale & Chall (1948; Chall & Dale 1995). Using simple features, mostly word length, word frequency and sentence length, these studies were able to predict large shares of variance in text comprehension scores. Later on, this work has been criticized by discourse psychologists for ignoring fundamental features of comprehension processes, especially the role of text coherence in these processes (e.g. Kintsch & Vipond 1979). The need was felt to automatically analyze coherence features as well. Eventually, this led to the development of Coh-

Metrix, an application that extracts both word- and sentence-based features and cohesion features from English text (Graesser et al. 2004). T-Scan shares these ambitions of Coh-Metrix, and adds one to them: the intention to support users in making text-specific diagnoses.

T-Scan will try to address text features that are deemed theoretically and practically relevant by human experts. Hence it is important that the T-Scan features are transparent: users will need to understand how feature values relate to specific textual characteristics, and how textual interventions affect feature scores. In this respect, T-Scan differs from text analytic applications that extract features from texts and classify them in order to save us the work of reading them. For instance, the latest generation of readability prediction tools uses statistical language modeling methods. Such methods compare the input text with language models (often lexically based) derived from texts deemed suitable for certain reader populations (e.g. Collins-Thompson & Callan 2005; for an overview, see Benjamin 2012). They train the tool to replicate target group assignments on the basis of the similarity between the text and particular language models. Such a tool may be useful in tasks such as choosing educational texts or tailoring search results to the level of the computer user. In contrast, T-Scan text features are not just means to an end, but deserve scrutiny in themselves, as they need to inform us about stylistic, structural and content characteristics that determine the complexity of texts.

This has several implications. First, document-level language models are a less attractive source of features, as they generate no transparent statements about text characteristics. Instead we need to develop intuitive and text specific features, and discuss their definitions in some detail, as they cannot be used unless understood. Second, we need to provide both text-level and local output for our features, so that results may point to particular text locations. Third, we need to be alert users to the possibility of faulty analyses, and consider whether they threaten the overall quality of the analysis.

Finally and crucially, we need to discuss the validity of our features, as many of them are quite shallow features in comparison to the insights we hope to glean from them. For instance, while a number of surface features may be used to assess how personal the writing is in a text, we need to know how far these measures are removed from the actual personal writing construct. Another conceptual issue is whether our features reflect writing style or text content. When only considering prediction performance, the style/content issue is not that important: anything that does the job is welcomed. For instance, strong lexical predictors such as word length may be seen as stylistic but may also be taken to suggest that content is more important than style: longer words generally go hand in hand with difficult, technical topics. Once we ask how different text characteristics affect processing ease and how text may be optimized, the style/content issue gets to be more pressing, as only stylistic features lend themselves to modification (given that a particular content is to be conveyed). Hence syntactic and cohesion features, which presumably vary more freely given text content, are just as interesting as lexical features, although they will probably not be equally strong comprehension predictors.

This is why this paper will introduce T-Scan not by showcasing its performance in prediction studies (but see Kraf & Pander Maat 2009 for a readability prediction study and Brummel 2013 for a genre prediction study using earlier T-Scan versions). Instead, we will provide a qualitative discussion of the tool. Using a small demonstration text corpus, we will review results for a number of central T-Scan features, raising three types of issues for each.

- Definition issues: how is the measure defined and computed?
- Reliability issues: what kinds of errors may occur in preprocessing and categorization?
- Validity and informativeness issues: to what extent does the measure reflect relevant aspects of the style, structure and content of the text?

This kind of reflection is needed before we may proceed to the next phase of our work: finding meaningful relations between T-Scan features on the one hand and text genres and text comprehension and processing data on the other hand.

2. T-Scan overview

T-Scan extracts text features in nine classes, illustrated in Table 1; more information is given in the T-Scan manual, available for users of the tool (Pander Maat et al. 2014).

Feature class	**Examples**
Lexical complexity	Word lengths in letters and (free) morphemes Word and lemma frequencies for two corpora Frequency rank class membership Nominalizations Prepositional expressions Abbreviations
Sentence complexity	Sentence length (Subordinate) clauses per sentence Passives Negations Dependency lengths (e.g. subject-verb, object-verb) NP length NP modifiers (number; kind)
Referential cohesion and lexical diversity	Type-token-ratio Measure of Lexical Diversity in Text Repeated arguments from sentence n-1 Repeated arguments from the last X words Anaphoric pronouns
Relational coherence	Connectives Spatial, temporal and causal words
Concreteness	Semantic type for nouns Semantic type for adjectives Semantic type for verbs Geographical, organization and product names
Personal style	Personal pronouns Personal nouns and names Action verbs
Verbs and time	Tense Aspect Action / process / state verbs
Parts-of-speech	Densities for 10 POS-tags
Probability features	Trigram probabilities Perplexities

Table 1: T-Scan feature overview

T-Scan draws on various tools and resources developed in the Dutch computational linguistics community in the last decades:

- Frog[1] (Van den Bosch et al., 2007): tokenization, lemmatization, PoS-tagging, named entity recognition;
- Alpino[2] (Bouma, Van Noord, and Malouf, 2001): dependency parsing;
- SoNaR[3] (Oostdijk et al. 2013) and SUBTLEX-NL[4] (Keuleers et al. 2010): frequency lists;

[1] http://ilk.uvt.nl/frog
[2] http://www.let.rug.nl/vannoord/alp/Alpino/
[3] http://tst-centrale.org/nl/producten/corpora/sonar-corpus/6-85
[4] http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-nl

and

- Referentie Bestand Nederlands[5] (Martin & Maks 2005): semantically annotated word lists.

3. A tour along some central T-Scan measures

3.1 Our example corpus

In this paper, we use an example corpus of six texts of around 300 words. Three texts come from the Dutch social science journal Mens en Maatschappij (further MenM) and the other three are taken from a weekly magazine for young women, Flair. The MenM texts are on interethnic stereotyping in inner city neighborhoods, differences in the social network of persons who have been in prison and persons who have not, and explanations for extremist political attitudes. Two of the three Flair texts are columns recounting personal experiences, one of them is about skin care. Appendix 1 provides the first MenM excerpt, Appendix 2 the first Flair column.

Obviously, Flair and MenM texts differ vastly in their topics, communicative purposes and target audiences. And in fact, even the most shallow features point to large differences (Table 2). Flair seems to offer less complex writing in that it has much shorter words than MenM, and often shorter sentences as well. Flair shows a larger syntactic variety, also both between and within texts (note the high ratio of standard deviations to means for the two Flair columns). In the course of this paper, we will try to deepen our insight in these differences: to what extent are MenM texts really more complex than Flair texts, and if yes, what underlies these complexity differences?

Text	**Word length (letters)**	**Sentence length (words)**
Flair 1 (column)	4.4 (0.9)	11.0 (6.2)
Flair 2 (skin care)	4.4 (0.7)	15.6 (4.2)
Flair 3 (column)	4.3 (0.4)	18.3 (11.0)
MenM 1 (ethnic relations)	6.1 (0.8)	23.8 (10.5)
MenM2 (ex-prisoners)	5.9 (0.9)	22.3 (7.2)
MenM3 (extremism)	6.0(0.6)	20.1 (7.8)

Table 2: Word lengths and sentence lengths in the example corpus

3.2 Lexical complexity

Different corpora, different word frequencies

Word length is an extremely crude indication of lexical complexity. The validity of word frequency indices is better already, since presumably word frequencies model the amount of exposure to word forms or lemmas, and hence may help predict word familiarity, i.e. the entrenchment of the form-meaning association (Just & Carpenter 1987). Breland (1996) provides some straightforward support for the link between word frequency and word knowledge. However, word frequencies are taken from frequency lists based on corpora, and different corpora may provide different perspectives on the lexical make-up of text.

In T-Scan, we use frequencies from SoNaR (Oostdijk et al. 2013) and SUBTLEX-NL (Keuleers et al. 2010). SoNaR is a written language corpus of more than 500 million words containing text from many genres and domains, although it needs to be stressed that more than half of it comes from newspapers. Subtlex-NL is a 44 million word corpus drawn from Dutch subtitles to American films and television series. It stands to reason that Subtlex-NL frequencies model a different language

[5] http://tst-centrale.org/nl/producten/lexica/referentiebestand-nederlands/7-20

register than SoNaR frequencies, as subtitles mainly contain scripted informal conversation. Its size makes it more attractive than an existing corpus of informal conversations, the Spoken Dutch Corpus, which is considerably smaller.

T-Scan provides two kinds of frequency features: logs of frequencies per million words and frequency rank classes. The second measure is most readily understandable. It simply tells us how many of the texts words belong to the first 1000, 2000, 3000, 5000, 10000 and 20000 words in the corpus. Table 3 provides frequency class proportions for our two texts, both those based on SoNaR and those based on Subtlex-NL.

	Top1000	**Top2000**	**Top3000**	**Top5000**	**Top10000**	**Top20000**
MenM1 - SoNaR	.59	.68	.74	.80	.86	.90
Flair1 SoNaR	.70	.78	.80	.81	.87	.90
MenM1 - Subtlex	.55	.60	.61	.69	.78	.84
Flair1 Subtlex	.73	.78	.80	.85	.89	.91

Table 3: Word frequency rank class proportions for two texts and two corpora

Unsurprisingly, the social science text has lower common words proportions than the Flair-text. The more interesting finding is that the difference between the two texts is larger for Subtlex frequencies than for SoNaR frequencies. In fact, the SoNaR difference disappears once we get past the Top5000. This suggests that Subtlex frequencies provide a more sensitive measure for distinguishing academic from everyday language registers.

This is confirmed by checking what words end up in different frequency classes in Subtlex and SoNaR (see Table 4). SoNaR has higher frequency ranks for words in the formal and academic registers, while Subtlex ranks are higher for words that prominently figure in everyday domestic and family interactions.

	Words whose SoNar frequency rank exceeds its Subtlex rank	**Words whose Subtlex frequency rank exceeds its SoNaR rank**
MenM	*contacten, context, culturele, ingegaan, nauwelijks, plaatsvinden, specifieke, theoretische*	*draagt, hiervan, relatie*
Flair	*achtergrond, gelegenheid, ruimschoots, samenvatting, vervolgde, zaterdagochtend*	*achterin, broek, cool, dozen, jasje, mannetje, pap, papa, stoere*

Table 4: Words differing in SoNaR and Subtlex frequency ranking

This already adds valuable information to the results in Table 2, and enhances the validity of our lexical diagnosis. The words in MenM are not just longer than in Flair, but less frequent, especially in everyday conversation.

Adding morphological analysis may improve lexical complexity predictors

A second, still unresolved issue in defining lexical features involves the status of compounds and derivations. Anderson & Davison (1988) have pointed out that word frequencies and word lengths may be misleading indicators of word difficulty: semantically transparent compounds and derivations are less frequent and longer, but not necessarily more difficult than their component or root words, except perhaps for beginning readers and other readers poor in decoding skills. Hence the T-Scan team is working on frequency measures that correct for the possibly misleadingly low frequencies of compounds. The issue is a pressing one for compounds, as many new words that return zero frequencies are compounds.

We consider using a frequency measure based not on the frequency for the compound word (e.g.

parkeergarage, contacthypothese) but on the autonomous frequencies of the free morphemes making up the compound (*contact + hypothese*). This requires us to draw on the morphological analysis provided by Frog, one of the central tools in the T-Scan machinery. Currently, Frog's morphological analyzer (MBMA) works on a high level of granularity (Van den Bosch and Daelemans, 1999), so that the free morphemes extracted will often not be those that are immediately apparent. For instance, gebruikersvriendelijk (user-friendly) will not be decomposed into [gebruiker] + [s] + [vriendelijk] but into [gebruik] + [er] + [s] + [vriend] + [elijk]. Now it might be doubted whether knowing the meaning of *friend* automatically implies knowing the concept of *friendliness*, and whether the metaphorical extension of *friendliness* in combinations such as *user-friendliness* is a transparent one for every reader. Hence it remains to be seen whether using the frequencies of free morphemes from this decomposition will provide us with more valid word difficulty indices.

3.3 Sentence complexity

Dependency lengths: strengths and weaknesses

Like word length, sentence length is another classic readability prediction feature. It can be measured quite reliably, but has been criticized for its lack of causal validity (e.g. Kemper et al. 1993). One of the candidates for replacing it is dependency length (Gibson 2005, Temperley 2007): not the length of a sentence is what taxes readers, but discontinuities in its syntactic structure. Drawing on analyses provided by the Alpino dependency parser (Bouma et al, 2001), T-Scan provides the distance for a number of dependencies, for instance subject-verb, verb-verb particle and determiner-noun. When a sentence contains several instances of a dependency type, T-Scan provides mean lengths for it. Let us consider some examples.

1. De contacthypothese is hevig bekritiseerd, met name omdat de theoretische noties van de contacthypothese te algemeen zijn en er nauwelijks wordt ingegaan op de specifieke culturele en sociaal-economische context waarin interetnische contacten plaatsvinden.

In sentence 1, the longest mean dependency (length: 4) is the one concerning the relation between subordinating conjunctions and the head verbs of their clauses. The mean of 4 is based on three dependencies of this type:

- *omdat - zijn* (length: 8 words)
- *en - ingegaan* (length: 2 words)
- *waarin - plaatsvinden* (length: 2 words)

Aside from the construction type means, T-Scan also provides the length of the longest dependency it encounters over all the dependencies of all types to be found in the sentence. In this sentence, the longest dependency is the first conjunction-head verb dependency. Hence when diagnosing a sentence the user should look whether it contains a really long dependency (say, beyond 7 words), and if yes, look for this particular dependency among the dependency types with relatively large mean distances.

It is important not to get distracted by single-instance dependencies. In the following sentence, there is just one relation between conjunction and dependent clause verb to be considered, namely the one between *terwijl* and *liggen* (length: 9 words):

2. Opvallend is dat de contacthypothese, in zoverre wij dat kunnen overzien, niet in verband wordt gebracht met studies die veiligheidsbeleving bestuderen in relatie tot 1) sociale cohesie/sociaal kapitaal en 2) het interetnische karakter van wijken, terwijl dat gezien onze literatuurbespreking voor de hand lijkt te liggen.

However, there is a longer dependency in the sentence: the subject-verb dependency between *contacthypothese* and *gebracht* (length: 11). This is not immediately obvious as the mean length for subject-verb dependencies is based on a number of subject-verb constructions as well: *contacthypothese - wordt (10), wij - kunnen (1), wij - overzien (2), studies - bestuderen (2), dat - lijkt (6) and dat - liggen (8).*

Now do dependencies really add information that is not already in the sentence length? Of course, there is a correlation between sentence lengths and dependency lengths such that longer sentences have longer dependencies. However, for the 103 sentences in our example corpus the correlation between sentence length (SL) and the longest dependency length (LDL) in the sentence is .69, which is substantial but far from perfect. When only taking the social science sentences into account, the correlation drops to .60. We checked this tendency for an educational text corpus of 100 excerpts, taken from 100 secondary education textbooks on four subjects: Geography, History, Dutch and Economics. The corpus contains 2677 sentences altogether (mean sentence length 12.0 words). The SL-LDL correlation between is .71. However, for the 1021 sentences longer than 12 words (mean sentence length 17.8 words), the correlation drops to .47. For longer sentences, the SL-LDL relation is considerably weaker; and it is here that we most need a reliable complexity indicator.

To appreciate the potential divergence between the two measures, compare the next two sentences, taken from the second Flair text.

3. Helemaal bijkomen en huid, haar, maar ook je body, mind and soul helemaal opladen.
4. Een lang warm dampend bad zorgt ervoor dat je lichaam volledig kan ontspannen, inclusief alle spieren.

The first sentence has 14 words but contains long dependencies, such as the one between the conjunction en and the verb head of its second conjunct (length 10); this verb is also quite far away from its first object noun *huid* (length 9). In contrast, the second sentence counts 17 words but has no dependencies longer than 3 words (e.g. *een bad*).

We need to realize that Alpino, despite being a robust parser, does make mistakes now and then. For example, it may have trouble correctly attaching prepositional phrases in long sentences. In the following sentence, the final phrase *in multiculturele wijken* modifies the immediately preceding noun, but is coupled instead to the main verb *zien*, giving rise to a spurious dependency spanning 15 words:

5. Als we de besproken onderzoeksvelden vergelijken, zien we dat een integratie daarvan een verrijking kan betekenen van de theoretische reflecties over veiligheidsbeleving *in multiculturele wijken.*

Such spuriously long dependency lengths compromise the diagnostic quality of our scan. We will need to do more testing on long sentences to learn more about the Alpinos reliability in these especially challenging contexts.

Clause length and adjectival modifiers

Other T-Scan syntactic complexity indices are simpler than dependency length but still might be more valid than sentence length, one of them being clause length. It has been suggested that sentence length is not taxing in itself as long as the sentences present information in clauses, as these are often related by conjunctions (e.g. Irwin & Pulver 1984, Cain & Nash 2011). Many longer sentences contain more than one finite clause: the correlation between sentence length and number of finite clauses is substantial in our example corpus (r=.65). Perhaps adding clauses to a sentence is less important to complexity than adding words to a simplex sentence or clause. Hence clause length may be a more valuable indicator of processing load than sentence length.

T-Scan determines finite clause length for each sentence by dividing the number of words by the number of finite verbs in the sentence. In passing we note that examining finite verbs is also interesting for another reason: it brings to light elliptical sentences without finite verbs. While these sentences are not taken into account for clause lengths, they are an interesting feature by themselves: besides signaling an informal style, they also affect sentence length. The short mean sentence length in the first Flair column is largely due to the five verbless clauses in this text.

Now let us see what clause length may have to do with complexity. The following sentence in the first MenM text has 25 words on just one finite verb:

6. Dit heeft enerzijds te maken met de concentratie van minderheden in een wijk en anderzijds met de concentratie van personen met een zwakke sociaal-economische positie.

Information is densely packed in this sentence, especially by means of adjective phrases: much of the propositions in the sentence (in the sense of Kintsch & Keenan 1973) are realized by building complex noun phrases. The climax of this style is reached in the next sentence, which contains nine adjectival modifiers, both adjectives and prepositional groups (underlined and numbered).

7. De lage^{1} sociale^{2} betrokkenheid vertaalt zich in een laag^{3} sociaal^{4} vertrouwen en een laag^{5} informeel ^{6} toezicht, wat overigens ook geldt voor homogene^{7} wijken met een concentratie^{8} van sociaal-economisch zwakkeren^{9}.

T-Scan measures the incidence of adjectival modifiers both in terms of their number per clause (for sentence 7 this number is 4.5: 9 (modifiers) / 2 (clauses)) and as a density, that is as a frequency per 1000 words (9 modifiers / 29 words gives a density of 310 modifiers for 1000 words).

When our first two texts are taken as a whole, adjectival modifiers are far more common in the social science excerpt than in the Flair column, as shown in Table 5. The table also shows that the difference varies according to the unit of computation. As the MenM sentences typically contain more than one finite clause whereas the Flair sentences do not, the sentence does not seem to be a relevant unit here. Comparing finite clauses for the incidence of adjectival modifiers might seem more sensible. Perhaps comparing noun phrases at the highest level in clause might be even better, but our present noun phrase counter does not filter out modifying noun phrases, so that comparing clauses is our best option when we want to stick to a natural linguistic unit. Table 5 shows that MenM clauses contain two times the number of adjectival modifiers of Flair clauses. As MenM clauses are longer than Flair clauses, the word density difference is a bit smaller though still considerable: 221 for MenM and 141 for Flair.

Text	**Sentence length**	**Adjectival modifiers / sentence**	**Clause length**	**Adjectival modifiers / clause**	**Adjectival modifiers / 1000 words**
Flair1	11.0	1.6	9.3	1.3	141.0
MenM1	23.8	5.3	11.9	2.6	221.0

Table 5: Different indices for adjectival modifier incidence

However we prefer to compute the adjectival modifier feature, the difference can be taken to mean that the social science texts referents are more complex entities, in the sense that they require more elaborate linguistic constructions to successfully refer to them. We conclude that T-Scan does offer syntactic complexity measures that offer better validity than sentence length. Further testing of the tool and further empirical work is needed to determine which of them are to be preferred.

3.4 Referential cohesion and lexical diversity

Generally, a text which often repeats words is easier to process than a text which keeps on adding new words. Even the simple intervention of repeating the same term for the same referent has been shown to improve reproduction and comprehension (Britton & Gülgöz 1991). A classic measure sensitive to word repetition is the type-token-ratio (TTR), calculated by dividing the number of different text words by the total number of text words. TTR is one of the four predictors in the most frequently used readability formula for Dutch: de CLIB (CITO Leesbaarheidsindex voor het Basisonderwijs; Staphorsius 1994). TTRs can be computed on the basis of word forms or lemmas; they can also be computed for only content words. This may be yield more information, as function words are repeated much more often than content words.

However, the TTR has been shown to be quite sensitive to text length, in the sense that further on in the text, new words appear at slower rates (Van Hout and Vermeer, 2007). This may hinder the comparison of texts of different lengths. As an alternative, Jarvis & McCarthy propose the Measure of Lexical Diversity in Text (MLTD), which is computed as the mean length of sequential word strings in a text that maintain a given TTR value (the default threshold value is .72, but T-Scan allows this value to be chosen by the user). To be more specific, MLTD evaluates the text sequentially for its TTR by tracking TTR after every word. For example: *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (.625) *people* (.556) and so forth. When the TTR factor passes the .72-threshold, the factor count increases by a value of 1, and the TTR evaluations are reset. In our example, MTLD would reset after *people* (.667) and continue by *for* (1.00) *the* (1.00) *people* (1.00). The text length in words is divided by the number of times the text reaches the chosen TTR level, in order to obtain the mean number of words required to do so (see Koizumi 2012 for more illustration). The higher this number, the longer it takes before words start to recur and the higher the lexical diversity.

We have discussed the calculation of MLTD to show that it is a locally oriented measure (text memory is continually emptied by the resets). It is still an open question whether readers have an equally short memory. Table 6 lists various TTR and MLTD measures for our example corpus.

Text	**TTR word**	**MLDT word**	**TTR lemma**	**MLDT lemma**	**TTR content words**	**MTLD content words**
Flair1 column	.60	107	.56	87	.82	226
Flair2 body care	.55	87	.52	82	.74	170
Flair3 column	.59	87	.54	82	.86	265
MenM1	.53	80	.50	60	.73	137
MenM2	.53	92	.49	87	.71	135
MenM3	.54	100	.51	88	.75	168

Table 6: Lexical diversity measures for the example corpus

The Flair columns are more lexically diverse than the social science texts, especially for content words. Interestingly, the second Flair text, which provides tips for skin care, shows values comparable to the social science articles. This suggests that diversity measures not just reflect lexical choices but also the structure of the text: it seems natural for narrative texts that keep introducing new events to have higher diversity values than expository or argumentative texts that elaborate some central point or topic. This last characteristic seems to apply to both the skin care and the social science texts.

Another assessment of the referential structure of the text is offered by argument overlap measures. Argument overlap refers to the degree to which a text segment repeats referential expressions from a preceding segment. The expressions assessed here are anaphoric and personal pronouns, nouns, names and main verbs. As T-Scan is unable to actually track referential chains through the text, we are confined to counting word form or lemma overlap. Hence we adopt the simplifying assumption that identical nouns, names and verbs, as well as pronouns of the same person and number (e.g. he/his) are share their referent. (This need not necessarily be so: for instance, a conversation in which subsequent utterances contain first-person pronouns is seen as high in argument overlap, although the pronouns will often not be coreferential.)

T-Scan offers two kinds of argument overlap: sentence overlap and buffer overlap. Sentence overlap examines links between adjacent sentences; this makes sense as sentences are natural units for discourse structures to unfold. On the other hand, differences in sentence length may compromise the comparison here. Hence the second overlap measure compares stretches of text of fixed lengths to be chosen by the user. In the default setting, the buffer size is 50 words. Roughly, this size reflects referential overlaps with the preceding four sentences. T-Scan starts by comparing word 51 for overlap with words 1-50, next comparing word 52 with 2-51, and so on. Every comparison yields an overlap density, and the densities are averaged to provide a text measure.

Text	**Sentence overlap**	**50-word overlap**	**Pronouns**
Flair1 column	57	126	165
Flair2 body care	51	126	152
Flair3 column	51	153	190
MenM1	49	85	81
MenM2	54	92	96
MenM3	53	87	109

Table 7: Argument overlap and pronoun densities for the example corpus

The second column in Table 7 shows no real differences in sentence overlap; but on the longer 50-word stretches the Flair texts seem to show more referential continuity.

In order to assess the validity of the overlap features, we manually annotated the first Flair column and the first MenM text for referential chains. For each chain member, we checked whether it was identified by our overlap features. This exercise yielded several observations. Evidently, the overlap measures miss anaphoric expressions that lack lexical overlap with their antecedent. This does not only hold for synonyms, but also for pronouns following nouns such as *daarvan* in example 8. Second, the sentence overlap measure misses overlaps within sentences, such as the link between we and we in example 8; buffer overlap measures do identify such cases. And third, the first MenM text contains quite a few long-distance anaphors that will only be captured when working with larger buffer sizes than 50 words. For instance, the concept (*multiculturele*) wijk is repeated three times in the text subsequent to example 8, but the third repetition is too far from the second to count as overlap.

8. Als we de besproken onderzoeksvelden vergelijken, zien we dat een integratie daarvan een verrijking kan betekenen van de theoretische reflecties over veiligheidsbeleving in multiculturele wijken.

The manual analysis also showed that the larger continuity in the first Flair text is partly due to personal pronouns which are repeated regularly across the text, but often not in the sentence that follows immediately; 57% of the pronoun overlaps in this text is only visible in the buffer overlap measure. In contrast, the MenM text has a much lower personal pronoun density. Personal pronouns are responsible for most of the pronominal overlap: other pronominal anaphors are typically not repeated. Table 7 shows that these observations generalize to the mini-corpus: personal pronoun densities differ markedly between the Flair and MenM texts, whereas there is no clear difference in the densities of other pronouns.

This is not to say that personal pronouns are a better and simpler way to capture referential relations than the buffer overlap feature. Only a combination of both measures can show us that the larger continuity in the second Flair text is not only due to personal pronouns, but to numerous repetitions of words such as *oil* and *skin* as well. This also helps to explain the lower lexical diversity of this text (see Table 6).

We conclude that cohesion and lexical diversity need to be analyzed together. When taking 50-word overlap as our cohesion measure and content word TTR or MLTD as lexical diversity feature, the following picture emerges. The two narrative Flair texts combine high referential continuity (largely due to pronoun chains) with high lexical diversity, whereas the other instructional Flair text shows high continuity based on both pronouns and content words, resulting in lower lexical diversity. The MenM texts show lower continuities combined with lower lexical diversities. This is partly due to the fact that many of their recurrences (mostly noun chains) span larger distances.

This means that two related dimensions are crucial in talking about referential structure: the distance between anaphor and antecedent as well as the kind of expressions dominating the referential chains (pronouns versus nouns). Both dimensions are well-known from theoretical work on referential coherence (from Ariel 1988 on). These dimensions provide insight in genre differences, which provides them with prima-facie validity. In future research we need to determine how they affect text complexity and whether they may help in comparing the complexity of texts within the same genre.

3.5 Lexical semantic features: a window on text content

The noun classification

It is a linguistic commonplace that the interpretation of words depends on their context; nonetheless, many words can still be categorized in terms of the entities, situations and properties they typically refer to. The use of such semantic word lists has been pioneered by the Linguistic Inquiry and Word Count (LIWC), which also has a Dutch version (Zijlstra et al. 2004). This version offers 6600 words classified into categories such as emotions, cognitive processes, perceptual processes, friends, family, school and job. For T-Scan, we aimed at a larger coverage and drew on the lists of nouns (36000 words), adjectives (9000) and verbs (6700) provided by the Referentie Bestand Nederlands (RBN; Martin & Maks 2005). The classifications in the RBN noun list were interesting in that they shed light on word concreteness. For a tool that aims to assess text complexity, word concreteness is more interesting than identifying topic domains such as family or school. Hence we set out to manually correct the RBN noun list. The RBN adjective list was partly recategorized to provide more direct concreteness information, whereas the verb list was extended with concreteness coding. More details on all classifications can be found in the T-Scan manual (Pander Maat et al. 2014).

The noun list contains 36850 lemmas classified into eleven categories, see Table 8. Polysemous words that could be read as belonging to several of the first eleven classes were left undefined.

Class	**Examples**
1. Human	leraar, schreeuwlelijk
2. Nonhuman	mus, eik
3. Artefact	stoel, weefgetouw
4. Substance	arsenicum, ijswater
5. Concrete, other	galblaas, vulkaan
6. Place	Amsterdam, voorkamer
7. Time	feestdag, periode
8. Measure	euro, dB
9. Dynamic abstract	angstkreet, loonverlaging
10. Nondynamic abstract	Christendom, hoogte
11. Institution	Werkgeversorganisatie
12. Undefined	poot, verband

Table 8: RBN noun semantic types used in T-Scan

The first five classes can be seen as concrete words in that they refer to humanly perceptible spatio-temporal entities, whereas the classes 9, 10 and 11 are most abstract. T-Scan looks up all words tagged as nouns, and outputs densities and proportions for al twelve classes. To illustrate what kind of information this may provide us with, Table 9 presents some noun classes that show differences in our example corpus. For example, it says that in the first Flair column 55 words were tagged as noun, of which 84% (46) were found in the list; furthermore, 30% of these 46 nouns referred to artefacts and 9% to substances.

A particularly strong genre cue in Table 9 is the proportion of nondynamic abstract nouns: social science articles are replete with such nouns (*onderzoek, invloed*), in contrast to Flair texts. To a lesser extent, social science texts stand out by their proportions of dynamic abstract nouns (event nouns) such as *verrijking* and *beeldvorming* are also characteristic for social science texts. The Flair articles have more strict concrete words referring to artefacts (*skateboard*), substances (*olie*) and other spatio-temporal entities (*huid*).

	n	**Listed**	**Human**	**Art.**	**Sub.**	**Conc. other**	**Place**	**Time**	**Dyn.**	**Non-dyn.**	**Un-defined**
Flair1 column	55	.84	.20	.30	.09	.02	.04	.07	.02	.09	.17
Flair2 skin care	68	.91	.00	.11	.16	.24	.02	.06	.06	.15	.18
Flair3 column	52	.96	.24	.14	.02	.12	.00	.12	.04	.14	.18
MenM1	62	.84	.04	.00	.00	.02	.12	.00	.13	.46	.23
MenM2	67	.93	.16	.00	.00	.00	.03	.03	.11	.48	.18
MenM3	60	.85	.18	.00	.00	.00	.02	.00	.08	.63	.10

Table 9: Noun class proportions for the example corpus (art=artefact, sub=substance, conc=concrete, dyn=dynamic)

Table 9 also shows how individual texts may stand out for their use of particular word classes. For instance, the first social science text talks a lot about neighborhoods, leading to a higher incidence for place nouns, whereas the second Flair text differs markedly from the other two, not only it its strategy of person references, but also in its emphasis on substances and parts of the human body (concrete other).

The first two columns in Table 9 show that our list does reasonably well in terms of covering the texts vocabularies. T-Scan outputs the words not yet covered, so that the lists can be periodically updated.

The verb classification

We have manually checked the RBN list of about 6600 verbs for which action, process and state readings are listed. As it happens, many verbs allow different readings in different contexts. For instance, *tekeergaan* may be an action by some person or may represent a process brought about by a nonhuman force (e.g. a storm). In our revised list, most polysemous verbs have been left undefined, except for the action/process cases, which have been labeled processes.

As a new verb feature, we have distinguished between concrete and abstract verbs. Concrete verbs represent situations or processes that can be directly perceived, e.g. seen or heard. Verbs which do not evoke sense experiences have been called abstract, while verbs that have both readings (depending on the context) have been left undefined for abstractness, see Table 10.

ASP reading / code	**Abstract**	**Concrete**	**Undefined**
Action	*aanbesteden, afgelasten*	*kwetteren, lassen*	*verfrissen, verlichten*
Process	*ineenstorten, meemaken*	*doorlekken, openrijten*	*leeglopen, losslaan*
State	*toeschijnen, hopen*	*vriezen, maffen*	*ontbranden*
Action / process: coded as process	*ontkrachten, tekeergaan*	*doorboren, kronkelen*	*breken, neerslaan*
Action / state: undefined	*beantwoorden, letten*	*hobbelen*	*paren*
Process / state: undefined	*frustreren, meevallen*	*ruiken*	*horen*
Action / process / state: undefined	*bijdragen, verschaffen*		*hechten, maken*

Table 10: The T-Scan verb classification (ASP=Action/State/Process) and verb concreteness codes

We will first present results for verb concreteness here; later on we return to action verbs. Although concrete verbs rarely occur (see Table 11), the pattern seems clear: they only occur in the womens magazine (e.g. *kleden, dampen, insmeren*). Two MenM texts stand out because of their abstractness proportions; for the third text, the picture is blurred by a number of undefined verbs that have abstract readings in this context, such as *zien* and *stellen* (e.g. de contexttheorie stelt).

	n	Listed	Concrete	Abstract	Undefined
Flair1 column	48	.96	.02	.74	.24
Flair2 skin care	42	.98	.05	.59	.37
Flair3 column	48	.96	.07	.70	.24
MenM1	40	1.00	.00	.65	.35
MenM2	54	.98	.00	.89	.11
MenM3	57	.98	.00	.91	.09

Table 11: Verb concreteness proportions for the example corpus

The adjective classification

The final source of lexical-semantic features is based on our adjectives list, which contains a new categorization of the RBN adjectives. Compared to the RBN scheme, the new taxonomy introduces additional distinctions between human and nonhuman attributes (categories 1-2 vs. 3-4), between kinds of observability (3 vs. 4), between evaluative and non-evaluative abstract words (7-11 vs. 12) and between several kinds of evaluation (7, 8, 9, 10 and 11). Table 12 provides the new taxonomy.

Class	Examples
1. Humanly observable characteristics of people	*doodsbleek, dwergachtig*
2. Emotional characteristics and social behavior	*gegriefd, goedgelovig*
3.Humanly observable characteristics of nonhuman entities	*druipnat, flanellen, geel*
4.Characteristics that are not humanly observable	*teerarm, kiemvrij*
5. Time	*voorbijgaand, vrijdags*
6. Place	*binnenlands,Gelders*
7. Specific evaluation (positive/negative)	*onverslijtbaar; lawaaierig*
8.General evaluation (positive/negative)	*mooi; verwerpelijk*
9.Epistemic evaluation (positive/negative)	*steekhoudend; onzinnig*
10.Evaluation by intensifiers	*ontzaglijk, ontembaar*
11.Evaluation by downtoners	*beperkt*
12.Abstract, non-evaluative	*aanverwant, aandachtig*
13.Undefined (adjectives with multiple readings)	*belastbaar, druk, small*

Table 12: T-Scan adjective types

The adjective data in Table 13 show how social science texts are dominated by abstract adjectives (*sociaal, direct, historisch*) whereas Flair texts more often use undefined everyday adjectives with large numbers of readings (*groot, lang, druk, los*) or evaluative adjectives *(goed, fijn, heftig*).

	n	Listed	Human / emotional & social	Observable nonhuman	Time	Evaluative classes	Abstract non-evaluative	Un-defined
Flair1 column	36	.83	.03	.00	.20	.20	.03	.53
Flair2 skin care	33	.97	.00	.03	.09	.41	.06	.41
Flair3 column	20	.90	.00	.11	.11	.39	.06	.33
MenM1	43	.91	.00	.00	.05	.23	.67	.05
MenM2	29	.93	.00	.00	.04	.11	.85	.00
MenM3	35	.94	.06	.00	.00	.24	.55	.15

Table 13: Adjective class proportions for the example corpus

Despite their obvious shortcomings given the pervasiveness of polysemy, our lexical-semantic data have deepened our insight in the semantic and pragmatic differences that are behind the word length differences in Table 1. We have already seen that Flair words are more frequent in everyday conversation. This section adds some semantic substance to these findings: Flair words are less often abstract, lend themselves to various concrete readings and are more often evaluative.

3.6 Personal style features

Ever since Flesch (1948), the presence of a human element has been regarded as potentially helping the text to communicate clearly. And Argamon et al. (2003) have shown that personal style markers differentiate between genres (fiction vs. non-fiction) and author gender. At present, T-Scan offers the following features to examine the presence of human attributes:

- Personal and possessive pronouns (1st, 2nd, 3rd person; total)
- Nouns referring to persons (one of the categories in the noun semantics list)
- Person names (based on the named entity recognition module in Frog)

T-Scan also adds up these three kinds of person references to a grand total.

Arguably, the proportion of verbs referring to actions is also relevant, as it indicates the presence of human actors in the text world. Let us see how these features behave in our example corpus.

	Pronoun 1st	Pronoun 2nd	Pronoun 3rd	Pronouns all	Human nouns	Person names	Total pers-ref	Action verbs
Flair1 column	74	10	30	114	30	0	145	.15
Flair2 skin care	3	61	3	68	0	0	68	.30
Flair3 column	84	29	23	135	39	13	186	.23
MenM1	24	0	10	34	10	7	51	.22
MenM2	10	0	22	32	32	3	67	.30
MenM3	0	0	46	46	30	17	93	.21

Table 14: Human elements in the example corpus (pers-ref= person references; all values are densities per 1000 words, except that action verbs is a proportion of verbs)

Obviously, the Flair texts use more pronouns than the MenM texts; in the columns, first-person pronouns predominate, while the skin care text directly addresses the reader. Third-person pronouns do not really differ. Human nouns do not differ either, as MenM texts use plural (and generic) human nouns such as *personen* and *extremisten*, while in Flair there is talk about singular (and referentially specific) terms such as *papa* and *zoon*. See Appendices 1 and 2 for illustrations.

As for person names, we have several problems. First, their identification is not yet very reliable. T-Scan misses person names such as *Ollie* and *Rokeach*, which are both seen as place names. The third MenM text refers to nine other authors, two names of whom are correctly labeled person names.

A more fundamental problem is that lists of author names that are not involved in semantic predications, do not actually signal human elements in text (see Appendix 1, sentence 2). These lists of names are a quite unique genre feature of scientific text and compromise the validity of person names as indices of a personal style. T-Scan users will need to be aware of such peculiarities.

Let us finally consider action verbs. It may come as a surprise that the first Flair column has less action verbs than the second MenM text, so let us inspect the verbs in both texts. In the first Flair column we find a number of undefined verbs referring to actions in this context (e.g. *aantrekken, komen*). The MenM text on the social network of (ex-)prison inhabitants contains a few past participles lemmatized to action verbs (e.g. *gedetineerden*), but most action verbs actually refer to actions, albeit abstract ones (e.g. *tijd besteden, relaties onderhouden*). Furthermore, the authors themselves may be actors (*wij baseren ons op*).

We conclude that, except for first- and second-person pronouns, the human elements as defined by T-Scan do not necessarily make a text more personal. If we want our personal style construct to focus on individuals that readers can identify with, it might help to further distinguish between singular and plural person references, even though singular references may still be generic ones and it is uncertain whether generic references can be filtered out automatically. Moreover, we need to further test the strengths and weaknesses of our person name recognition engine.

4. Conclusions

4.1 What does T-Scan tell us about texts?

What have our explorations taught us about our example corpus, and about our instrument? Let us start with the first question. T-Scan offers a number of lexical measures that bring to light the vast differences in text content in our small text collection. The word frequency profiles, especially those based on Subtlex, show how our first two texts tap into different domains of experience. The lexical-semantic data for nouns and adjectives show enormous differences between the proportions of abstract words on the one hand, and sizable differences between the proportion of concrete nouns and evaluative adjectives on the other. That is, our various corpora frequencies and lexical semantic data differentiate between various Flair genres in a much more informative way than is possible with word length differences.

Likewise, personal and possessive pronouns are a simple but non-trivial indication of the primary protagonist in text; especially first and second person pronouns differentiate clearly between genres.

All this primarily concerns text content. What can T-Scan tell us about writing styles, about the choice of linguistic forms to convey particular concepts? We have discussed a number of syntactic complexity indices that may be smarter than sentence length: dependency lengths, clause lengths, adjectival modifiers. We will need to further test these syntactic indices, both in corpus work and in our upcoming prediction studies, as they enable us to better diagnose and optimize sentence structures.

Referential cohesion and lexical variation may well be predictive of genre and comprehension differences, but we need to further examine how particular referential structures translate into TTR, MLTD and argument overlap scores, and how the contribution of referential structure (what is referred to?) interacts with that of lexical variation (which referential expression is chosen?). Another issue is how local and global measures may differ.

4.2 Future work in T-Scan development

A continuing source of future work is reducing errors in categorization and preprocessing. Processing steps such as lemmatization are involved in many T-Scan measures, and may introduce errors by themselves. These imperfections come as no surprise for computational linguists, but may surprise T-Scan users. For one thing, we will have to get used to part-of-speech tags that follow the Spoken Dutch Corpus annotation guidelines on phrasal verbs that are realized in separate parts: for instance, the verb *opladen* is seen as a separate *op* and *laden* in the sentence ”Dat geeft het lijf de kans om de batterij weer op te laden”. This choice leads to inflated word frequencies and incorrect lexical-semantic coding.

Another issue caused by the adherence to the Spoken Dutch Corpus annotation guidelines for part-of-speech tagging is that present and past participles are coded as verbs, including those modifying nouns (*de verworpen verzoekschriften; zoekende ogen*) or functioning as nouns (*de verworpenen der aarde; de zoekenden*). However, subsuming these cases under the generic verb tag is not optimal for T-Scan purposes, as these differently positioned participles function differently in the semantic representation of sentences. Hence we will need to create the option to correct our verb density for (pre)nominal participles, which can presumably be done by using the deeper morpho-syntactic layers of Frogs POS-information.

These are just a few of the short-term tasks facing us; we have already mentioned the need to correct compound frequencies. In the long run, we hope to include semantic similarity information into T-Scan, which might help us to obtain deeper kinds of cohesion measures that do not capitalize on surface repetitions.

The lesson of this project so far is that building the application and analyzing texts need to go hand in hand, as features need to be monitored for reliability and validity as well as for the most relevant unit of computation. Every feature offers a particular trade-off between reliability on the one hand and validity and informativeness on the other. This trade-off determines its usability for various analytical purposes. Hence computational and discourse linguists will need to cooperate closely in further optimizing T-Scan.

References

Anderson, R.C. & A. Davison (1988). Conceptual and empirical bases of readability formulas. In: A. Davison & G.M. Green, *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, NJ: Lawrence Erlbaum, pp. 23-53.

Argamon, S., Koppel, M., Fine, J. & Shimoni, A.R. (2003). Gender, genre and writing style in formal written texts. *Text* 23(3), 321-346.

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics* 24, 65-87.

Benjamin, R.G. (2012). Reconstructing readability: recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review* 24, 63-88.

Biber, D. & Conrad, S. (2009). *Genre, register and style.* Cambridge University Press, Cambridge.

Bouma, G., Van Noord, G., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers,* 37(1), 45-59.

Breland, H.M. (1996). Word frequency and word difficulty: a comparison of counts in four corpora. *Psychological Science* 7(2), 96-99.

Britton, B.K. & S. Gülgöz (1991). Using Kintschs computational model to improve instructional text: effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology* 83(3), 329-345.

Brummel, G. (2013). The usability of T-Scan for automatic genre classification. Internship paper UiL-OTS, Utrecht.

Cain, K. & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology* 103(2), 429-441.

Chall, J. S. & Dale, E. (1995). *Readability revisited: The new DaleChall readability formula.* Cambridge, MA: Brookline Books

Dale, E. & J.S. Chall (1948). A formula for predicting readability. *Educational Research Bulletin* 27, 37-54.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology.* 32(3), 221-233.

Gibson, E. (2000). The Dependency Locality Theory: a distance based theory of linguistic complexity. In In Y. Miyashita, A. P. Marantz & W. ONeil (eds.), *Image, language, brain* Cambridge: MIT Press, 95126.

Graesser, A. C., McNamara, D. D., Louwerse, M. L., & Cai, Z. (2004). Coh-Metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*36(2), 193202.

Just, M.A. & P. Carpenter (1987). *The psychology of reading and language comprehension.* Allyn & Bacon, Newton Massachusets.

Kemper, S., Jackson, J.D, Cheung, H. & Anagnopoulos, C.A. (1993). Enhancing older adults reading comprehension. *Discourse Processes* 16(4), 405-428.

Irwin, J.W. & Pulver, C. J. (1984). Effects of explicitness, clause order and reversibility on childrens comprehension of causal relationships. *Journal of Educational Psychology* 76(3), 399-407.

Kintsch, W. & J. Keenan (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology* 5(3), 257-274.

Kintsch, W. & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In LG. Nilsson (Ed.) *Perspectives on Memory Research* (pp. 329-366). Hillsdale, New Jersey: Erlbaum.

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods,* 42(3), 643-650.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction* 1(1), 60-69.

Kraf, R. & Pander Maat, H. (2009). Leesbaarheidsonderzoek: oude problemen en nieuwe kansen. *Tijdschrift voor Taalbeheersing* 31(2), 97-123. (Readability research: old problems and new opportunities)

Martin, W. & Maks, I. (2005). Referentie Bestand Nederlands. Met medewerking van S. Bopp en M. Groot.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381-392.

Oostdijk, N., Reynaert, M. Hoste, V. & Heuvel, H. van den (2013). *SoNaR User Documentation.* Version 1.0.4.

Pander Maat, H., Kraf, R., Dekker, N., Sloot, K. van der, Bosch, A. van den, Gompel, M. van & Kleijn, S.(2014). *Handleiding T-Scan.* Available at webservices-lst.science.ru.nl.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.* Cito.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition* 105 (2), 300-333.

Van den Bosch, A., and Daelemans, W. (1999). *Memory-based morphological analysis. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 285-292.

Van den Bosch, A., Busser, G.J., Canisius, S., and Daelemans, W. (2007). *An efficient memory-based morphosyntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde (Eds.), Computational Linguistics in the Netherlands 2006: Selected Papers of the Seventeenth CLIN Meeting.* Utrecht: LOT, 191-206.

Van Hout, R., and Vermeer, A. (2007). Comparing measures of lexical richness. In Daller, H., Milton, J., Treffers-Dahl, J, et al. (Eds.), *Modelling and assessing vocabulary knowledge.* Cambridge, UK: Cambridge University Press, 92-115.

Zijlstra, H. Meerveld, T. van, Middendorp, H. van, Pennebaker, J.W. & Geenen, R. (2004). De Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC), een gecomputeriseerd tekstanalyseprogramma. *Gedrag en gezondheid* 32, 271-281.

Appendix 1

Excerpt 1 from Mens en Maatschappij

1. De contacthypothese is hevig bekritiseerd, met name omdat de theoretische noties van de contacthypothese te algemeen zijn en er nauwelijks wordt ingegaan op de specifieke culturele en sociaal-economische context waarin interetnische contacten plaatsvinden.
2. Toch blijft de contacthypothese aantrekkelijk voor wetenschappers en wordt zij gebruikt in recent onderzoek naar stereotypering van bijvoorbeeld moslims (Novotny & Polonsky, 2011, Savelkoul, Scheepers, Tolsma & Hagendoorn, 2010).
3. Opvallend is dat de contacthypothese, in zoverre wij dat kunnen overzien, niet in verband wordt gebracht met studies die veiligheidsbeleving bestuderen in relatie tot 1) sociale cohesie/sociaal kapitaal en 2) het interetnische karakter van wijken, terwijl dat gezien onze literatuurbespreking voor de hand lijkt te liggen.
4. Hieronder beschrijven we de conceptuele samenhang van de eerder besproken studies, waarbij we nader ingaan op de relatie tussen sociale cohesie, interetnische interacties, beeldvorming en veiligheidsbeleving.
5. Als we de besproken onderzoeksvelden vergelijken, zien we dat een integratie daarvan een verrijking kan betekenen van de theoretische reflecties over veiligheidsbeleving in multiculturele wijken.
6. Als we de studies op elkaar afstemmen in een samenhangend geheel, dan ziet dat er als volgt uit.
7. In een multiculturele wijk is de sociale betrokkenheid beperkt.
8. Dit heeft enerzijds te maken met de concentratie van minderheden in een wijk en anderzijds met de concentratie van personen met een zwakke sociaal-economische positie.
9. De lage sociale betrokkenheid vertaalt zich in een laag sociaal vertrouwen en een laag informeel toezicht, wat overigens ook geldt voor homogene wijken met een concentratie van sociaal-economisch zwakkeren.
10. Wat een extra dimensie geeft aan sociale betrokkenheid in een multiculturele wijk, is het interetnische karakter hiervan.
11. Als die beperkt is, leidt dit tot negatieve beeldvorming.
12. Deze beeldvorming draagt vervolgens direct bij aan de beleving van onveiligheid in de buurt en indirect omdat zij een negatief effect heeft op sociaal vertrouwen en informeel toezicht.

Appendix 2

Excerpt 1 from Flair

1. Kinderen worden zo snel groot.
2. Kun je ze niet af en toe sealen zodat ze niet ouder worden?
3. 'Pap, ik wil een skateboard voor mijn verjaardag.'
4. 'Goed zoon, gaan we doen.'
5. Ziehier, een summiere samenvatting van een goed vader-zoon gesprek.
6. 'Toen ik tien was, maakten we zelf een skateboard met een plank en vier wieltjes van een gevonden winkelkar van de Aldi.
7. Die wieltjes lieten te snel los, ' vervolgde ik.
8. 'Pap, ik weet waar we moeten zijn.'
9. Het belooft een mannendag te worden.
10. En dus sta ik op zaterdagochtend in een skateboard afgezakte broeken winkel met mijn mannetje.
11. De muziek is wat heftig en net iets te hard voor op de achtergrond, zou ik zo denken.
12. Ja, papa wordt oud, ik weet het.

13. Mijn zoontje wordt ook ouder, want hij is ook al 'bijna' tien en hij vindt het hier helemaal 'top'.
14. Grote zoekende ogen want hij is best onder de indruk maar zijn stoere slenter-sleep loopje compenseert dat ruimschoots.
15. 'Goeiemiddag, ik kom voor een skateboard voor mijn zoontje.'
16. 'Ok man, cool, Ollie kan je ff helpen?'
17. Ollie blijkt een opgeschoten lange jongen van een jaar of twintig te zijn.
18. Achterin de winkel staat hij dozen in of uit te pakken, dat is mij niet helemaal duidelijk.
19. Maar hij is er best druk mee.
20. Lage zwarte broek, grote gympen met losse veters, een zwart T-shirt met een cartoon gothic kop erop en op zijn eigen hoofd een (hoe kan het ook anders) zwart mutsje.
21. 'Hey man, alles flex ?' ...
22. 'Euh ja zeker, alles kits achter de rits... '
23. Ik voel mijn zoontje duizend doden sterven.
24. Ik kijk naar Ollie en Ollie kijkt naar mij.
25. Ik heb voor de gelegenheid mijn favoriete jasje aangetrokken.
26. Camelkleurig van wol met een fijn visgraatje.
27. Niets leukers dan je als ouwe lul te kleden.