Der Test zur Erfassung der standarddeutschen Vorleseaussprache lief zumeist reibungslos ab. Nur von wenigen Gewährspersonen wurde dieser Erhebungsteil als eine Art Prüfungssituation empfunden, sodass eine Unsicherheit entstand. Der zweite Vorlesetext (Zeitungstext), der einige lange Linksattribute, Namen und Fremdwörter enthält, wurde von manchen Gewährspersonen als schwierig und zu lang empfunden; hier sind bei der Datenanalyse eventuelle Einflüsse auf die Artikulation im zweiten Teil des Textes zu bedenken.

**5.5. Datenaufbereitung** Die Aufnahmen wurden mithilfe des im SFB 538 „Mehrsprachigkeit" an der Universität Hamburg entwickelten Transkriptions- und Annotationsprogramms EXMARaLDA („Extensible Markup Language for Discourse Annotation"), einem System von Konzepten, XML-basierten Datenformaten und Werkzeugen für die computergestützte Transkription und Annotation sowie das Erstellen und Auswerten und die Onlinevisualisierung von Korpora gesprochener Sprache transkribiert und für die sprachwissenschaftliche Bearbeitung aufbereitet (Kellner et al. 2008). Die Kernkomponenten des Systems sind: (1) ein Partitur-Editor, der für die Transkription (inkl. Alignierung mit einem Audiosignal) und Annotation der Aufnahmen in einer Partiturdarstellung eingesetzt wurde, (2) das Suchwerkzeug Exakt („EXMARaLDA Analyse- und Konkordanztool"), mit dem die linguistische Annotation des Korpus durchgeführt wurde, und (3) das Korpusverwaltungswerkzeug Coma („Corpus Manager"), mit dem Metadaten zu Sprechern und Erhebungssituationen verwaltet wurden. Genauere Angaben zur technischen Verarbeitung und nachhaltigen Sicherung der im SiN-Korpus gewonnenen Sprachdaten finden sich in dem Projektband Schröder i. Vorb.

Aus den Aufnahmen der Tischgespräche wurden als Grundlage für die Variablenanalyse Stichproben aus den Redebeiträgen der Hauptgewährspersonen gezogen. Für die Stichproben wurde bewusst keine gezielte Auswahl bestimmter Gesprächspassagen vorgenommen. Vielmehr wurden ab einem bestimmten Zeitpunkt (meist ab Minute 10) die Redebeiträge der Gewährsperson (unter Ausklammerung der Beiträge der anderen Gesprächsteilnehmer) solange transkribiert, bis ein Umfang von etwa 2500 Wörtern erreicht war. Hierdurch wurden alltägliche, auch banale Kommunikationsakte erfasst, u.U. auch mit länger andauernden Gesprächspausen, Unterbrechungen usw. Aufgrund der selektiven Verschriftlichung enthalten die transkribierten Stichproben keine kohärenten Gespräche, sondern nur Hilfstranskripte, die als Referenzspur für die Annotation dienen. (Gesprächsanalytische Transkripte nach GAT 2 wurden für ausgewählte Passagen im Rahmen des Teilprojekts 3 am Standort Münster angefertigt.) Um das Korpus mit den verfügbaren Suchroutinen erschließen zu können, folgte die Verschriftlichung der standarddeutschen Orthographie (ohne Interpunktionszeichen), standarddivergente Lautrealisierungen wurden auf dieser Stufe noch nicht vermerkt, sondern erst später annotiert. Verbale Äußerungsformen wie Gesprächswörter, Rückversicherungspartikeln, Abtönungspartikeln usw. wurden dabei berücksichtigt, Verzögerungsmarker (*äh, öh*), paraverbale Merkmale oder nonverbale Aktivitäten blieben hingegen ausgeklammert. Eindeutige Fälle von Code-Switching ins Niederdeutsche wurden markiert und blieben bei der nachfolgenden Variablenanalyse unberücksichtigt. Längere niederdeutsche Redeanteile wurden gesondert erfasst.

Bei den Redeanteilen der Gewährspersonen gab es starke Divergenzen, sie reichten jedoch in der Regel


Abb. 2. Verteilung von hochdeutsch basierten (blau) und niederdeutsch basierten (rot) Stichproben im Korpus der Tischgespräche (weiß = keine Stichproben)

aus, um auswertbare Stichproben zu ziehen. Bei 101 Tischgesprächen, die ausschließlich oder überwiegend in hochdeutsch basierten Varietäten geführt worden waren, betrug der Umfang der Stichproben jeweils ca. 2500 Wörter. Bei weiteren 21 Tischgesprächen, die partiell in niederdeutscher Sprache geführt wurden, konnten hochdeutsche Gesprächsanteile im Umfang von mindestens 500 Wörtern berücksichtigt werden. Insgesamt besteht das Korpus der hochdeutsch basierten Stichproben aus den Tischgesprächen damit aus 122 Stichproben, die im Durchschnitt einen Umfang von ca. 1500 Wörtern haben. Nur in 22 Tischgesprächen waren die hochdeutschen Gesprächsanteile der Gewährsperson zu gering, um ausgewertet werden zu können (weniger als 500 Wörter). Abb. 2 stellt die Verteilung von hochdeutsch basierten (blaue Kästchen) und niederdeutsch basierten (rote Kästchen) Stich- proben im Korpus dar. In zwei Fällen liegen für beide Sprachen keine ausreichenden Stichproben der Gewährsperson vor (weiße Kästchen). Genauere Informationen zu den nieder deutsch basierten Stichproben aus den Tischgesprächen bietet der zweite Atlasband.